NAVAL HEALTH RESEARCH CENTER

STATISTICS AND THE ART OF CONSTRUCTION

R. R. Vickers, Jr.

Report No. 04-08

Approved for public release; distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (M2)
2300 E ST. NW
WASHINGTON, DC 20372-5300

STATISTICS AND THE ART OF MODEL CONSTRUCTION

Ross R. Vickers, Jr.

Human Performance Department
Naval Health Research Center
P.O. Box 85122
San Diego, CA 82186-5122
e-mail: Vickers@nhrc.navy.mil

Report Number 04-08, supported by the U.S. Army Medical Research and Materiel Command, Ft. Detrick, Frederick, MD, under Work Unit No. 60109. The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution unlimited.

This research has been conducted in compliance with all applicable Federal Regulations governing the protection of human subjects. No human subjects were directly involved in this research.

Introduction

Constructing and evaluating behavioral science models is a complex process. Decisions must be made about which variables to include, which variables are related to each other, the functional forms of the relationships, and so on. The last 10 years have seen a substantial extension of the range of statistical tools available for use in the construction process. The progress in tool development has been accompanied by the publication of handbooks that introduce the methods in general terms (Arminger, Clogg, & Sobel, 1995; Tinsley & Brown, 2000a). Each chapter in these handbooks cites a wide range of books and articles on specific analysis topics.

Recent developments are too broad to cover in detail in a single chapter. Instead, this chapter examines current statistical modeling practices from a particular perspective: How can available statistical tools be translated into practices that improve the quality of behavioral science models?

Framing the Problem

For the purposes of this chapter, a high quality model is defined as a model that is grounded in reliable knowledge. Reliable knowledge is produced when a scientific community reaches a consensus interpretation (Ziman, 1978), developed through a process of principled argument (Abelson, 1995). The acronym "MAGIC" summarizes critical elements of this process. M, *magnitude*, is "… the quantitative support for the qualitative claim" made by a theory. A, *articulation*, is "… the degree of comprehensible detail in which conclusions are phrased." G, *generality*, is "… the breadth of applicability of the conclusions." I, *interestingness*, is hard to define, but Abelson (1995) suggests "… to be *theoretically* interesting it must have the potential, through empirical analysis, to change what people believe about an important issue" (italics in the original). C, *credibility*, is the believability of a research claim. Credibility depends on methodological soundness and the theoretical coherence of the claim.

Statistical methods contribute to reliable knowledge when they promote principled argument. This chapter examines trends in data analysis methods from that perspective. First, the chapter will consider construction of measurement and substantive models. These models are considered separately because different methods (e.g., factor analysis vs. regression) traditionally have been used for the two purposes. The second section of the chapter covers methods of model appraisal and amendment. This section begins with a brief review of debates related to significance testing and then considers methods of augmenting this practice by using effect sizes (ESs), confidence intervals (CIs), and goodness-of-fit indices (GFIs). Issues related to defining and choosing between alternative models (e.g., confirmation bias) are also discussed. Finally, the chapter examines qualitative

analysis and exploratory data analysis as methods that can be applied to extend existing models.

Overview of Model Construction Methods

The following examination of available model construction methods is organized around three themes. The first theme is a distinction between measurement models and path models. This element of the discussion emphasizes the fact that most data analysis problems can be assigned to one of two broad categories. Some problems involve construct measurement. Other problems involve determining the relationships between different constructs. Methods in the first category produce measurement models. Methods in the second category produce substantive models. Substantive models describe causal paths linking different constructs. For this reason, these models are sometimes referred to as path models. Table 1 illustrates these categories with reference to specific methods covered in this discussion.

The intended endpoint of most behavioral research is a causal model. A progression from measurement models to explanatory models is natural in this context. For example, early research on Gulf War syndrome and posttraumatic stress disorder included substantial efforts to demonstrate that these syndromes existed. Initial efforts developed measurement models to represent each syndrome. This problem had to be solved before moving on to the development of path models. Path models were then constructed to identify causes and effects of these illnesses. This development pattern is a normal sequence that supports consideration of measurement and path models as separate but related topics.

The second underlying theme of this discussion focuses on the nature of the constructs under study. Meehl (1990b) captures a fundamental dichotomy by characterizing variables as "differences in degree" or "differences in kind." Constructs that involve differences in degree (e.g., temperature) can take on any value between the high and low anchor points of the metric scale. Many personality variables (e.g., neuroticism) and organizational climate measures (e.g., group cohesion) involve differences in degree. In contrast, constructs that involve differences in kind classify entities into qualitatively distinct groups. Gender and psychiatric classifications are common examples of this type of construct. Scales that measure differences in degree are referred to as continuous measures; scales that measure differences in kind are referred to as discrete or categorical measures. Different analysis models are appropriate for each class of variable.

Table 1. Model Construction Methods

Measurement Models

Dimensional Models

Exploratory
- Exploratory factor analysis (EFA)
- Multidimensional scaling (MDS)

Confirmatory
- Confirmatory factor analysis (CFA)

Categorical Models

Exploratory
- Exploratory cluster analysis (ECA)
- Latent class analysis (LCA)

Confirmatory
- Expectation-Maximization Mixture Analysis
- Taxometrics

Path Models

Dimensional Models

Exploratory
- Regression, including multiple regression
- Analysis of variance (ANOVA)
- Hierarchical Linear Models (HLM)*

Confirmatory
- Structural equation modeling*

Categorical Models

Exploratory
- Categorical and limited dependent variables (CLDV)**

Confirmatory
- Taxometrics
- Latent class analysis (LCA)

*Includes latent growth curve analysis (LGCA) as a special case.
**Includes logistic regression, logit analysis, probit analysis, loglinear analysis, and other specific models as special cases.

Finally, the discussion of analytic methods will draw a distinction between exploratory and confirmatory analyses. This theme focuses on the constraints imposed on a model. Pure exploratory analysis imposes the minimum number of constraints required to perform an analysis. Pure confirmatory analysis completely constrains the model by specifying the theoretical constructs that are involved, linking each observed variable explicitly to one or more constructs, and specifying the parameter values for the functions that convert the observed values into estimates of the theoretical constructs. Factor

loadings and regression coefficients are familiar examples of the parameters specified within confirmatory models. The pure forms of exploratory and confirmatory analysis can be regarded as endpoints of a continuum. The typical analysis imposes weak constraints. For example, the number of factors in a factor analysis may be fixed without constraining which items define what factors, or how large the individual factor loadings will be. The exploratory/confirmatory dichotomy is a reminder of the endpoints on the continuum. Movement from the exploratory extreme toward the confirmatory extreme defines progress within a research domain.

Model Construction

Model construction has theory validation as its ultimate objective. This view provides a framework for covering the full range of issues associated with model construction. Theories involve claims about causal patterns. These claims are not required when the goal is simply to predict an outcome. Prediction only requires a statistical association between a criterion and one or more predictors; associations do not necessarily have to indicate causal relationships. A predictive model can be evaluated simply by determining whether it is accurate when it is applied to new samples from the original population (i.e., cross-validation) or to samples from different populations (i.e., generalizability).

Theory validation imposes additional criteria. The associations among variables must be both consistent with the theory being tested and inconsistent with competing theories. Also, causal inferences must be justified.

Measurement models are required for testing theories. From this perspective, the measurement problem is to determine whether a coherent measure is formed by the observed patterns of association among hypothesized indicators of a construct. It is necessary to demonstrate that this basic assumption is justified before testing any hypothesis about the relationship between the measured construct and other constructs. For example, the antecedents and consequences of depression cannot be studied until depression itself can be measured.

Given a measurement scale, researchers can begin to investigate the pattern of associations between the measured construct and measures of other relevant constructs. The empirical pattern of associations determines scale validity (American Psychological Association [APA], 1985) even as it tests theoretical predictions concerning the pattern of associations among constructs. In this case, the predicted pattern of associations defines a substantive model. When several theories address the same topic, several substantive models have to be compared with the observed pattern of associations. The comparison provides the basis for identifying the most reasonable models in light of the available empirical evidence.

The full process of model construction requires the development of both measurement and substantive models. The methods used to construct measurement and substantive models are somewhat different. Some methods, such as multiple regression and factor analysis, are familiar to most researchers. Other methods, such as taxometrics and categorical and limited dependent variable (CLDV) analyses, are less familiar. It is critical that researchers understand the range of options that are available to them. This understanding will permit the investigator to mold the analysis to his or her research concerns. As a general principle, research issues should drive the researcher's choice of analysis methods, rather than vice versa. (For additional discussion of problem-driven vs. methods-driven research, see Ness & Tepe, this volume.)

Measurement Models

Differences in Degree

Most behavioral constructs are conceived of as differences in degree, and their measurement scales are typically developed using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Gorsuch's (1983) readable introduction provides guidance on all of the important elements of EFA. Fabrigar, Wegener, MacCallum, and Strahan (1999) also provide a summary of current EFA practices and recommendations for improving those practices.

A trend toward the use of CFA instead of EFA is one sign of movement toward stronger methods in behavioral research. The increasingly frequent use of CFA has been accompanied by publication of a number of texts that provide general introductions to the method. However, introductory texts often omit important topics (Steiger, 2001). Bollen's (1989) text remains a recommended choice for an introduction to these methods. Boomsma (2000) and McDonald and Ho (2002) provide similar recommendations with respect to CFA.

The decision to employ EFA or CFA is influenced by the prior research that is available to guide model development. As suggested by its name, EFA is the most appropriate analysis option when little is known about the structure of a domain. For example, a questionnaire constructed to measure leadership might consist of a large number of questions. Investigators could reasonably ask whether various types of behavior described by the questionnaire items effectively combine to define a single overall leadership style. Alternative models would divide the items into subsets that define two or more distinct leadership styles. EFA can be used to compare the unidimensional and multidimensional model alternatives.

Exploratory Factor Analysis (EFA) in Practice. Fabrigar et al. (1999) have summarized current EFA practices from the perspective of three basic decision points in a factor analysis. First, a method of factoring must be chosen. Second, the number of factors to extract must be determined. Third, a criterion for rotating the factors to a

final solution must be chosen. The analyst has a number of options at each decision point. The most common choices are principle components analysis (PCA) with Kaiser's (1960) criterion to select the number of factors, followed by an orthogonal varimax rotation (Fabrigar et al., 1999). Computer programs probably contribute significantly to this apparent preference. In many statistical packages, this analysis results when an investigator simply provides a set of variables for analysis and accepts the default values for each program option. Investigators therefore may "choose" PCA without realizing they have made a choice.

The decision regarding how many factors to extract is important for theory formulation and testing. This decision defines the number of theoretical constructs to be represented in the measurement model. Obviously, this decision is critically important in modeling. Thus, it is noteworthy that Kaiser's (1960) criterion (i.e., $\lambda \geq 1.00$) often extracts *too many* factors. The magnitude of this problem increases when more variables are analyzed and/or when the sample size is small (Buja & Eyuboglu, 1992; Cota, Longman, Holden, Fekken, & Xinaris, 1993; Lautenschlager, 1989). Parallel analysis (Horn, 1965; Humphreys & Montanelli, 1975) or Velicer's (1976) minimum average partial procedure provide better guidance. These procedures have not been used much in the past because they have been difficult to implement. Recently published computer routines can solve this problem (Kaufman & Dunlap, 2000; O'Connor, 2000). Guidelines based on Monte Carlo studies can also be used to reduce the risk of extracting too many factors (Buja & Eyuboglu, 1992; Cota et al., 1993; Lautenschlager, 1989). The newer methods promote parsimony by reducing the risk of constructing models that include phantom theoretical constructs that are defined by chance patterns of covariation within a particular database.

Extracting too many factors may not be a major problem for EFA. Overextraction has little effect on the structure of true factors (Wood, Tataryn, & Gorsuch, 1996). The greater problem lies in wasted effort devoted to speculating about the meaning of what are essentially chance findings. Because subjectively plausible interpretations can be devised for factor analyses of random data (Armstrong & Soelberg, 1968), interpretability is not a good guide for factor retention. Decisions based on sample size and number of indicators are also of limited value in avoiding overextraction (MacCallum, Widaman, Shang, & Hong, 1999). Factor structure is more strongly determined by the *quality* of the indicator variables. When EFA is truly exploratory, it may be difficult to ensure that selected indicator variables meet this criterion.

Fabrigar et al. (1999) concluded their review of the state of the art with a set of recommendations that would improve on the typical current practice. Principal factors analysis should replace PCA. Oblique rotation (i.e., correlated factors) should replace orthogonal rotation. Parallel analysis or a direct measure of fit between the

factor model and the fit of the model to the data (e.g., root mean square error of approximation [RMSEA]) should replace Kaiser's criterion. Applying these methods to reanalyze data from three studies, Fabrigar et al. (1999, p. 291) concluded that "… an EFA with an oblique rotation provides much better simple structure, more interpretable results, and more theoretically plausible representations of the data than a PCA with an orthogonal rotation."

In the long run, getting the number of factors right in EFA is important, but not essential. Factors that are the product of chance associations will not reproduce in other samples. When somewhat similar chance factors are found in different studies, these factors will either have no relationship to other variables or will show different patterns of association across studies. Either outcome should discourage investigators from taking those factors seriously when constructing theoretical statements. Extraneous factors will thus be weeded out in cumulative research programs. Nevertheless, the research process is made more efficient when supported by intelligently focused analyses. The scientific progress is slowed by researchers who adopt the complacent view that the evidence will "sort itself out" in the long run.

Multidimensional Scaling (MDS). MDS is another method of establishing the dimensionality of a measurement space. This method can be applied to the same correlation matrices used in EFA. However, the interpretation of the elements of the matrix is different. In EFA, correlations indicate shared causal influences. Viewing EFA as a path model, the correlations between indicator variables can be estimated by multiplying the factor loadings (i.e., the path coefficients; cf. Kenny, 1979). Factor analysis procedures produce loadings that reproduce the observed correlations as well as possible within specific constraints.

MDS focuses on representing similarity between cases rather than common causal influences. MDS solutions define locations within a reference space. Cases that have similar attributes are close together; cases that are dissimilar are farther apart. This distance perspective can be implemented using a number of alternative methods to scale distance. No matter which distance measure is chosen, the analysis focuses on differences between the dimensional values rather than on the product of dimensional values (Davison, 1985; Davison & Sireci, 2000). The basis for setting the number of dimensions is the variance (in the observed distances) that is explained by adding another dimension to the model. The model is complete when the addition of another dimension would not substantially increase the variance explained. Compared with EFA, MDS typically requires fewer dimensions to describe relationships among entities (Davison, 1985).

Confirmatory Factor Analysis (CFA). EFA can be carried out with little or no knowledge of the likely structure of the data to be analyzed. Advance specification of the number of dimensions is optional in EFA. EFA computes loadings for all items on all

dimensions. In EFA, all dimensions are either orthogonal (uncorrelated) or oblique (correlated).

CFA places a greater burden on the investigator. The number and nature of dimensions in the model must be specified in advance. CFA requires at least an informed guess to support each of the three basic factor analysis decisions. The investigator must specify the number of latent traits to measure, designate which indicator variables define each latent trait, and specify a pattern of correlations between latent traits. The pattern can include a mixture of orthogonal and oblique dimensions. The model can be specified in greater detail by assigning specific values to factor loadings and factor correlations (cf., for example, Arbuckle & Wothke, 1999; Joreskog & Sorbom, 1981).

Figure 1 illustrates a CFA model that consists of hypothetical negative affect and extraversion constructs. Ovals represent the hypothesized latent traits. Rectangles represent measured variables. Arrows indicate hypothesized causal effects of latent traits on measured variables. The numbers next to the arrows are the equivalent of CFA factor loadings and indicate the estimated strength of the causal effect. The two-headed arrow indicates a correlation between latent traits. The number associated with the arc indicates that the correlation is moderate in magnitude.

In the past, CFA required the analyst to define parameter patterns for a number of matrices. Today, CFA is much more accessible. This is underscored by the fact that the model in Figure 1 was constructed simply by drawing the picture and then linking the picture's components to variables in the database. The figure was constructed using two different computer packages (LISREL and Amos) to compare results. Most other commercial packages share this simple method of model construction. Simplified methods have certainly supported the more frequent use of CFA in behavioral research. However, to optimize use of this method, underlying measurement issues must be clearly understood.

Figure 1 illustrates several points. First, each measured variable receives an arrow from just one of the two latent traits. The other latent trait might exert an effect on each of the indicators, but the paths for these possible effects have been fixed explicitly at zero (i.e., no effect). This is an example of how CFA imposes a model constraint. If the same model were developed using EFA, the omitted arrows would appear as potential causal paths. EFA would estimate a model in which all possible latent trait-measured variable arrows are included.

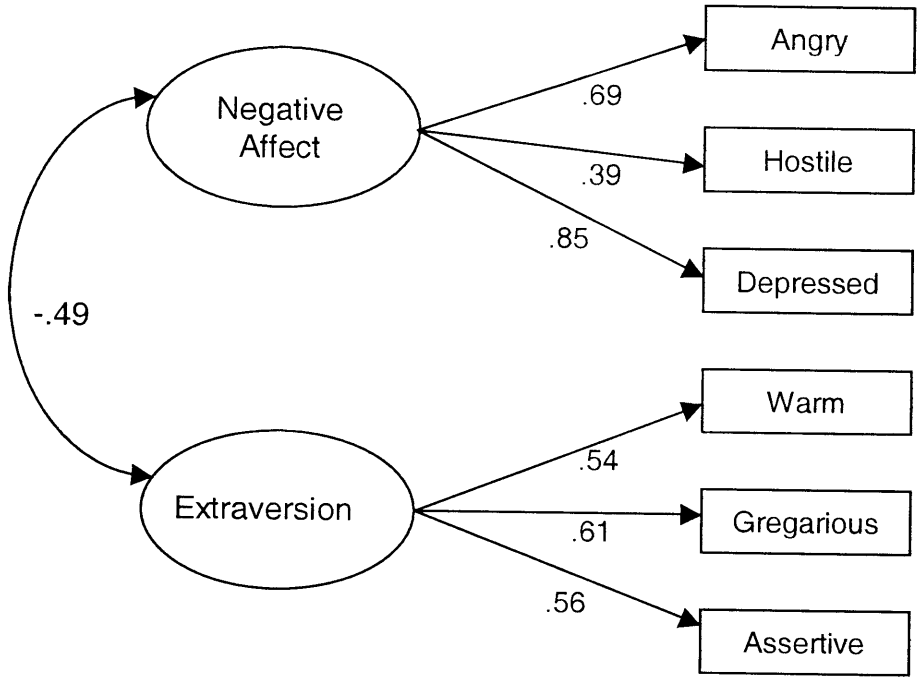

Figure 1
A CFA Model for Two Personality Traits

The second point illustrated by Figure 1 concerns the correlation between latent traits. This correlation is moderately large. An alternative model could have been defined with the correlation fixed at zero. The model in Figure 1 corresponds to an oblique rotation in EFA; the alternative model would treat neuroticism and extraversion as orthogonal factors. In this case, CFA removed a constraint that is imposed by the typical default EFA. CFA can test the plausibility of this constraint by comparing the goodness of fit of the orthogonal and oblique models. Methods for this comparison are described in the section on model appraisal and amendment.

The correlation between latent traits also can be used to illustrate the flexibility of CFA. The correlation in Figure 1 could be replaced by a causal effect. For example, an arrow from negative affect to extraversion would indicate a model in which negative affect caused people to be less willing to interact with others. Routine use of EFA would make the factors orthogonal (i.e., zero correlation). The correlation between latent traits could be introduced in EFA by using an oblique rotation, but EFA cannot impose a constraint that one latent trait is the cause of another. Causal interplay among personality variables is not a common topic of investigation, but it might be useful in some areas. For example, patterns of perception and

behavior that typify neurotic styles may have an underlying causal pattern (Shapiro, 1965).

A third point of interest regarding Figure 1 requires some amplification. The model is incomplete. A complete model would include a disturbance for each measured variable. A disturbance term is needed to reflect the fact that no measured variable is perfectly correlated with the associated latent trait. Imperfect relationships are expected, given that the measurement process generates some random error variance. The disturbance term combines random error and systematic variance from variables that are not in the model (James, Mulaik, & Brett, 1982).

CFA obviously requires more thought and effort than EFA. The payoff is greater flexibility in model construction. CFA provides stronger tests of models. CFA programs estimate parameters (i.e., factor loadings, factor correlations) subject to the specific set of constraints imposed by the investigator. If the resulting model accurately reproduces the observed pattern of covariation between indicator variables, the model has passed a riskier test than that embodied by EFA. The test is riskier because the EFA can adapt the number of dimensions, the pattern of factor loadings, and the pattern of factor correlations to the data. By contrast, the advance specification of these attributes of the model required in CFA increases the chance that the model will fail to reproduce the observed pattern of covariation. Thus, CFA is a relatively more demanding test. If the hypothesized model does account for the data, the result should be interpreted as stronger support (Meehl, 1990a).

CFA also provides tools for directly comparing alternative theories. This aspect of CFA is relevant when theories differ with regard to the range of behaviors relevant to different constructs in the model. In such cases, different theories will specify different numbers of dimensions and/or patterns of factor loadings for a given set of indicator variables. These differences specify different CFA models. The competing models can then be fitted to the data to determine which one does the best job of reproducing the observed patterns of correlation or covariance.

Models developed to deal with specific measurement issues illustrate the flexibility and power of CFA. CFA models have been developed to systematically quantify Campbell and Fiske's (1959) conceptualization of convergent and discriminant validity (Lance, Nobel, & Scullen, 2002; Marsh, 1989; Marsh & Bailey, 1991; Widaman, 1985). CFA can test circumplex models (e.g., Rounds & Tracey, 1993; Tracey & Rounds, 1993). CFA can test hypotheses derived from Lord and Novick's (1968) parallel/tau-equivalent/congeneric test classification (e.g., Millsap & Everson, 1991).

CFA also provides methods of addressing other recurring themes in the measurement literature. The generality of measurement models has been a longstanding concern (Blalock, 1982). CFA can evaluate

generality within a single study by fitting one model to the data gathered from two or more groups (e.g., males and females). Separate models then can be fitted for each group. If fitting separate models for each group does not substantially improve the overall fit, the first model applies to all groups. CFA can also be used to test the generality of previously published models even though the raw data from the model is not available for analysis. Factor loadings from a prior analysis can be used to define the model to be fitted to data from a new sample. Good fit is analogous to cross-validating a regression equation (Browne, 2000).

CFA can justify the use of standard psychometric models (e.g., internal consistency estimates of reliability). These models apply when indicator variables are effects of the underlying trait; these models do not apply when the indicators are causes of the trait (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Effect indicators are correlated because they share the latent trait as a common cause. Bollen and Ting (2000) provide a method of determining whether potential indicators are cause, effect, or a mixture of the two. This procedure may help clarify the structure of arguments over the proper interpretation of a scale.

Factor Analysis and the Accumulation of Knowledge. Reliable knowledge should be cumulative. From this perspective, EFA and CFA can play complementary roles. EFA is most useful when little is known about a measurement domain. Early EFA studies provide information that can be used to formulate theoretical statements that are the basis for later CFA studies. CFA studies can be used to develop and evaluate a sequence of models. The sequence can begin with relatively unconstrained models and move toward models that specify factor loadings and factor correlations.

Differences in Kind

Methods of assessing differences in kind have received relatively less attention than methods of assessing differences in degree. The greater emphasis on dimensional constructs is not necessarily an indication that these constructs are more important than categorical constructs. Meehl (1992) points out that whether a construct is dimensional or categorical is not a matter of choice or preference. The appropriate characterization is an empirical issue. From this perspective, the preference for dimensional measurements may be a form of bias.

Theoretical discussions that invoke typological language do not always indicate that a typological model is appropriate or intended. Sometimes behavioral researchers use typological language to simplify their communication. For example, psychologists may contrast extraverts and introverts to illustrate the extremes of a personality dimension. In other cases, typological language refers to real categorical entities. Psychiatric classification is probably the best

known example of this usage, and illustrates the potential for tension between typological and dimensional models. Some theorists argue for a dimensional model in this domain as well. These arguments echo Meehl's (1992) assertion that the choice between alternatives is an empirical issue. Even though categorical language should not always be taken at face value, there is a clear need for sound methods of defining categorical variables in empirical terms. Good methods are absolutely necessary for resolving measurement issues in some important behavioral domains.

Military researchers may be interested in typological variables in a variety of research contexts. Screening programs provide one obvious example. It can be wasteful and potentially harmful to train recruits who have personality disorders. Such individuals may leave the military before training costs have been recovered through their service. The stress of training and military life may produce lasting psychological difficulties for the trainee who already suffers psychological disorder. Sound typological measures may help to avoid such negative outcomes.

The exploratory-confirmatory continuum used above to contrast EFA and CFA can also be applied to differences in kind. The continuum is most evident in methods of cluster analysis. These methods traditionally have been applied with few restrictions on the structure of the resulting typology. Recent advances provide methods of imposing restrictions based on prior research. In addition, taxometric techniques provide alternative methods that focus on the existence of specific types, such as those found in psychiatric classifications.

Cluster Analysis. Cluster analysis is the most widely used method of defining categorical variables empirically. Cluster analysis begins with a set of cases (e.g., people, organizational units), each of which is represented in the data by a score profile. The analysis problem to define a set of groups that meet two basic criteria. First, cases assigned to the same group must have similar profiles. Second, the profile for the average case must differ substantially between groups. The first criterion can be satisfied best by having a large number of groups. Having many groups makes it possible for all cases within the group to have identical or nearly identical profiles. Having fewer groups most often satisfies the second criterion. This criterion provides a basis for combining groups with very similar, but not identical, profiles.

The data analyst must make several decisions when conducting a cluster analysis. The analyst must select a method (e.g., hierarchical agglomerative or direct), a similarity index (e.g., Euclidean distance with or without standardizing indicator scores), and a rule for determining group membership (e.g., average linkage or nearest neighbor linkage). Thus, cluster analysis is similar to factor analysis in the sense that the general procedure can take the form of a variety of different analyses for any given data set.

Different choices may lead to different results, so researchers must recognize the implications of the choices they make or read about in research reports. Gore (2000) provides an overview of cluster analysis. Aldenderfer and Blashfield (1985) provide a brief introductory text. Everitt, Landau, and Leese (2001) provide more detailed guidance, including advice about computer programs.

The results of exploratory cluster analyses (ECAs) must be viewed with skepticism. These procedures *always* produce clusters. The problem is not to decide whether there are clusters, but rather to determine how many empirical clusters represent real and distinct classes in the population. When Milligan and Cooper (1986) reviewed indices designed to help make this decision, the Hubert and Arabie (1985) adjustment to the Rand (1971) index was determined to be the best option. However, no single index is accepted today as the best option for determining the number of clusters (Gore, 2000). One reason may be that the rules are not available in standard computer packages.

The indices for establishing the number of clusters are based on the analysis of a single set of data. Replication of different cluster solutions across samples is another viable method. This method can be applied even in a single set of observations by dividing the set into subsets (Overall & Magee, 1992), but results must be interpreted carefully (Krieger & Green, 1999).

Empirically defined cluster solutions require further study to validate the typology. No matter which rule is applied, there is no guarantee that the resulting clusters will have any real meaning. The initial cluster definition can be a starting point for subsequent validation of the typology, but construct validation is a necessity before the typology is accepted as meaningful.

Recent developments provide a stronger approach to cluster analysis (Fraley & Raftery, 2002). Newer procedures are similar to CFA in structure because they provide the investigator the opportunity to constrain the cluster solution.

Recent developments in cluster analysis extend Wolfe's (1970) early work on mixtures of normal distributions. Today, the investigator can specify the structure of a group by defining the means, variances, and covariances for the indicators. The number of group structures specified determines the number of clusters in the analysis. Individual cases are assigned to clusters by computing the probability of membership in each of the hypothesized groups. The expectation-maximization (EM) algorithm (cf., McLachlan & Krishnan, 1997) is applied to compute these probabilities. The goodness of fit of the model to the data is indicated by χ^2 statistics produced by the EM algorithm, so significance tests are an option for determining the number of groups.

The EM approach to cluster analysis is a clustering analogue of CFA. Older procedures relied on mathematical criteria to define the number of groups and their boundaries. EM uses parameter values established by the analyst to define the groups. Those parameter values can be based on theory or past research. Constraints can be imposed to explicitly test alternative hypotheses about the structure of the typology. In the future, a distinction between ECA and confirmatory cluster analysis may become as common as the distinction between EFA and CFA.

Latent Class Analysis (LCA). LCA also defines typologies, and is used when the indicator variables being analyzed are categorical. In such cases, the data define a matrix with many different cells. Each cell represents one of the possible score profiles for the cases in the sample. LCA capitalizes on the fact that cases are not likely to be evenly distributed across all possible cells. Classes are subsets of the data set that represent particular score profiles. The profiles may be defined by the entire set of indicators or by a subset of those indicators.

As an example of the LCA problem, consider a disease that produces eight symptoms. If a questionnaire were constructed to determine who had the disease, each person who completed the questionnaire would respond "Yes" for each symptom that was present and "No" for each symptom that was absent. Taken together, the eight questions define 256 possible symptom combinations. However, suppose that four of the disease symptoms were physical and four were psychological. Some people might report suffering no symptoms at all, some only physical symptoms, some only psychological symptoms, and some might report having all eight (physical and psychological) symptoms. If these were the only response patterns that were observed, a four-group classification would be a reasonable representation of the data. The analysis problem would be to identify these four latent classes of respondents.

LCA tests for the existence of clustering in situations such as that described above. McCutcheon (1987) described the basic procedures and noted explicitly that LCA categories do not necessarily define a dimension. Conditional independence is fundamental to LCA. Independence is indicated by the absence of a correlation between responses on different indicators. Conditional independence has two elements. Independence is conditional if the indicators are uncorrelated within the classes in the LCA, but are correlated in the overall sample. LCA groups observations such that the resulting classification approaches conditional independence as closely as possible. The sum of the within-class χ^2 values is an index of how closely this objective has been approximated.

Several points should be kept in mind concerning LCA. First, a χ^2 significance test is the usual basis for choosing the appropriate number of classes. As discussed later in this chapter (see Model Evaluation), this criterion may tend to produce too many groups when

large samples are studied. With large samples, even small changes in the absolute fit of the model to the data will be statistically significant. Second, LCA suffers from indeterminacy problems. McCutcheon (1987, p. 25) notes that multiple solutions may provide equivalent fit to the data. Recent developments extend traditional LCA. Magidson and Vermunt (2002) distinguish LCA from latent class factor analysis (LCFA). LCFA defines categorical equivalents of EFA factors. Specifically, each LCFA dimension defines a categorical variable with two or more levels. Magidson and Vermunt (2001) demonstrate that the LCFA models can be more parsimonious than LCA models when considered at the level of the overall model. Thus, as a simple example, three parameters are needed to represent four groups. Only two parameters are needed to define the same four groups based on two dichotomous LCFs. Greater parsimony is indicated by the fact that fewer parameters are needed to correctly classify individuals (Popper, 1959).

Magidson and Vermunt (2002) have also shown that in at least some instances, LCA can classify cases more accurately than K-means clustering, which is the more common direct partitioning method of cluster analysis. In discussing their findings, Magidson and Vermunt (2002) note two important advantages of LCA over K-means analysis. The first is that classification involves computing the probability that each case is a member of each group. This probability then is used to weight the case when computing the group centroid, which avoids biased estimates that may be derived by the K-means approach of weighting all cases equally. Second, the LCA approach provides diagnostic information that can be used to determine the number of clusters. These χ^2 statistics can be used to construct GFIs. These indices, which are discussed in the Model Appraisal section of this chapter, provide a means of judging the accuracy with which a given model reproduces the observed distribution of cases across the cells in the cross-classification.

Taxometrics. Taxometric procedures (Meehl, 1992, 1995; Waller & Meehl, 1998) are another method of testing the claim that distinct groups exist within a population. This method focuses explicitly on a two-group solution. The hypothesized groups are a target group to be identified (i.e., taxon) and all others. Indicators of taxon status are assumed to be uncorrelated within each group (i.e., locally independent). Between-group differences produce correlations between the indicators in the general population. Taxometric procedures test for the existence of the hypothesized taxon and the base rates for that group and its complement.

The taxometric approach was developed initially in the context of a model of schizotypy, but this technique subsequently has been applied to other concepts such as "Type A" behavior (Strube, 1989). A recent simulation by Beauchaine and Beauchaine (2002) compared taxometric procedures to K-means cluster analysis. Taxometrics, specifically the maximum covariance procedure, was more effective "…when the number of indicators was few, when effect sizes were

reduced, when nuisance correlations were high, and when base rates were low" (p. 256). This comparison is limited by the use of a single taxometric procedure; convergence of several methods is preferred (Meehl, 1992, 1995). On the other hand, the use of EM-based clustering methods (Fraley & Raftery, 2002) also might have changed the results. This possibility is supported by a recent simulation study that found taxometric procedures and latent mixture modeling superior to direct cluster analysis (Cleland, Rothschild, & Haslam, 2000). However, the use of a poor criterion for deciding the number of clusters may have influenced these results.

Typological Analysis and the Accumulation of Knowledge. Cumulative knowledge is evident when a research domain moves from exploratory analysis to confirmatory analysis. Recent developments in typological analysis procedures provide tools to move the identification of "differences in kind" toward confirmatory modeling. Traditional cluster analysis and LCA procedures are largely exploratory since they provide only limited opportunities for the data analyst to constrain solutions. Both EM cluster analysis and taxometrics provide greater opportunities for using prior knowledge to constrain the models. LCFA analyses are primarily exploratory, but they are noteworthy because they provide alternative models that can be contrasted with LCA results.

The previous comments on factor analysis and knowledge accumulation apply to the construction of typological measurement models. Confirmatory models require parameter specifications. Confirmatory models represent cumulative knowledge when the parameter values have been derived from prior research. A priori parameter specification implies the same risk found in confirmatory tests of dimensional models. Results consistent with the model in spite of the risk provide stronger support for the model than the relatively qualitative evaluations provided by exploratory methods. The strategy of combining confirmatory methods and exploratory methods is good practice. This approach combines consistency (i.e., similarity to prior findings) testing with an exploratory search for better alternatives.

Measurement MAGIC

The preceding discussion of methods of developing measurement models has emphasized movement from exploratory to confirmatory models. Progress from the initial unstructured exploratory analysis to a final, completely specified confirmatory model should be accompanied by a general increase in Abelson's (1995) MAGIC criteria. Each element of MAGIC is briefly considered here.

The *magnitude* component of MAGIC conceivably could decline as a field of study progresses from early exploratory measurement models to final confirmatory models. Early models can account for more of the patterning in covariation or similarity measures by capitalizing on chance. The a priori specification of parameter values eliminates the

opportunity to incorporate chance into final models. Final models can be expected to exclude minor systematic sources of covariance and similarity (MacCallum, 2003). A final model will be acceptable if it captures the major systematic sources of variation in the indicator variables even if the overall explanatory power of this model is less than the apparent explanatory power that exploratory models achieve by capitalizing on chance.

The *articulation* of measurement models clearly increases with the progression from exploratory models to confirmatory models. A graph such as that in Figure 1 provides one way of thinking about model articulation. The model in Figure 1 is relatively simple because the number of latent traits is restricted and because some potential trait-indicator relationships have been ruled out. This simplification is articulation in the sense that it specifies particular traits and separates potential causal paths into those included in the model and those that either do not exist or are small enough to be ignored. For example, previous research by Costa and McCrae (1992) and others could have been used to specify the factor loadings for the model in Figure 1. Had this been done, the model would have been fully specified in advance of the analysis. The fully specified model would have provided complete articulation. Every latent trait, every causal effect, and every parameter value would have been specified a priori. If one compared a graph of the a priori specifications for EFA and CFA, the a priori definition for Figure 1 in EFA would consist solely of the list of variables at the sides of the figure. There would be no ovals or arrows. The additional specification of latent traits and causal effects in Figure 1 is a pictorial manifestation of model articulation.

Generality may or may not increase in the progression from exploratory to confirmatory models. While generality is usually assumed for new constructs, this attribute should be empirically established in the process of developing the measurement model (Blalock, 1982). CFA provides tests for generality that may extend to categorical analyses. The CFA methods for assessing generality rely on goodness of fit evaluations that could be readily adapted to the mixture approach to cluster analysis. However, tests for generality may indicate that different models are needed for different populations. If so, models should reflect that fact rather than treating different groups as equivalent.

Interest may be a constant in the sequence. Initial results often are interesting because they are novel and shed light on previously unknown territory. This interest should be tempered by an appreciation of the possibility that the findings are the product of chance. This interest also should be tempered by the realization that the results probably rest on assumptions that need to be tested. After initial model development, the clarity with which an analysis contrasts competing models may determine its interest value (Dixon & O'Reilly, 1999). As models mature, verification that a final model fits the data from a new sample should be interesting. However, at

this point, the strongest evidence might be the a priori prediction of the parameter values (i.e., the pattern of factor loadings) for a new indicator. This a priori prediction would be the type of "darned strange coincidence" (Salmon, 1984) that provides support for a theoretical model. Given these suggestions, interest in a field will be maintained by novel predictions and/or progression in articulating and contrasting alternative models.

Credibility should be greater for models that are near the confirmatory end of the model development sequence. This exploration should consider a range of plausible interpretations and alternative models for the constructs of interest and for mappings of indicators onto constructs. Credibility is enhanced by the use of confirmatory techniques to determine whether the indicators are effect variables, to rule out interpretations based on methods variance, and to demonstrate convergent and discriminant validity. The key to credibility is the model's cumulative track record (Meehl, 1990a). The final model should be the one that best accounts for the data from multiple studies.

Path Models

Path models describe relationships among theoretical constructs (McDonald & Ho, 2002). The methods available to construct path models include regression and analysis of variance (ANOVA). These procedures represent a subset of the possible methods for constructing and testing path models. The extensive range of alternatives available 20 years ago (Andrews, Klem, Davidson, O'Malley, & Rodgers, 1981) has grown (Andrews et al., 1998). This section will consider a shift in the rationale for choosing among alternative methods and will then examine structural equation modeling (SEM) and hierarchical linear modeling (HLM) as procedures that are likely to be used with increasing frequency in the future. Latent growth curve analysis (LGCA) and the analysis of CLDV are also considered as methods that simplify the approach to key modeling issues. These methods can be implemented in the context of SEM or HLM.

Selecting a Modeling Method

Twenty years ago, level of measurement issues might have been a primary consideration when selecting an analysis procedure (Andrews et al., 1981). The phrase "level of measurement" refers to the information content of the variables used in analyses. According to Nunnally and Bernstein (1994), nominal measures assign observations to categories that have no intrinsic ordering (e.g., male, female). Ordinal variables indicate relative magnitude (e.g., greater than, less than) for an attribute. Interval measures indicate order and the distance between observations. Ratio measures order observations, indicate distance between observations, and indicate distance from a zero value.

Level of measurement can be a guide to selecting analysis procedures. For example, Pearson product moment correlations would be computed to describe relationships between interval measures, but Spearman's rank order correlation would be computed to describe relationships between ordinal measures. This approach to choosing an appropriate statistic can be extended to a wide range of analysis problems. Doing so produces a complex decision tree with branches that often culminate in little known statistics (Andrews et al., 1981). Today, this approach is made easier by statistical software routines that make once obscure statistics more readily available.

Current trends shift the emphasis from data characteristics to the nature of the construct being studied. If the constructs being studied are continuous variables, analyses are chosen to obtain the best estimate of the population correlation that can be derived from the data. This focus is maintained regardless of the level of measurement. For example, consider the problem of estimating the relationship between two continuous constructs when one has been measured with an ordinal scale and the other with an interval scale. The level of measurement might lead to the use of a rank order correlation because rank order information is the least common denominator for the two measures. However, if both constructs theoretically are differences in degree, a polyserial correlation is an appropriate estimate of the true population correlation. Some standard analysis packages now would provide this theoretically appropriate estimate as a default (Joreskog & Sorbom, 1996). In effect, the information provided by the scales is used to obtain the best possible estimate of the theoretically relevant population parameter regardless of the level of measurement.

Structural Equation Modeling (SEM)

SEM techniques have been so widely applied that a detailed review of their uses would be unmanageable in any brief format. Bentler and Dudgeon (1996) and MacCallum and Austin (2000) provide recent overviews of these methods. Investigators who are just beginning to use SEM can choose from a growing number of texts. However, the choice should be made carefully. Many introductory texts gloss over critical considerations (Steiger, 2001). Bollen (1989) provides a sound introduction to the general method. The current chapter will limit consideration of SEM to specific recent developments that are particularly relevant to the range and content of models constructed using this tool.

Modeling Interactions. Interactions occur when the relationship between two variables is contingent on one or more other variables. For example, an investigator might be interested in determining whether the relationship between general intelligence and job performance was the same in different occupations. If the physical demands of different occupations were known, the question might be whether intelligence has less effect on performance in physically demanding jobs.

The use of SEM to model interactions is analogous to more familiar procedures. Analysis of covariance (ANCOVA) is one method of investigating interactions. Nonparallel regression lines illustrate the essence of an interaction. Consider a two-group analysis. In this case, a significant result in a test for nonparallelism of regression lines means $b_{11} \neq b_{12}$ (b_{1j} is the slope of a regression line in the j^{th} group). Moderated regression (Saunders, 1956) is another method. In this case, $y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$. The contingency in this case can be illustrated by considering that the slope of the regression for x_0 is b_0 when $x_2 = 0$, $b_1 + b_3$ when $x_2 = 1$, and $b_1 + 2b_3$ when $x_2 = 2$.

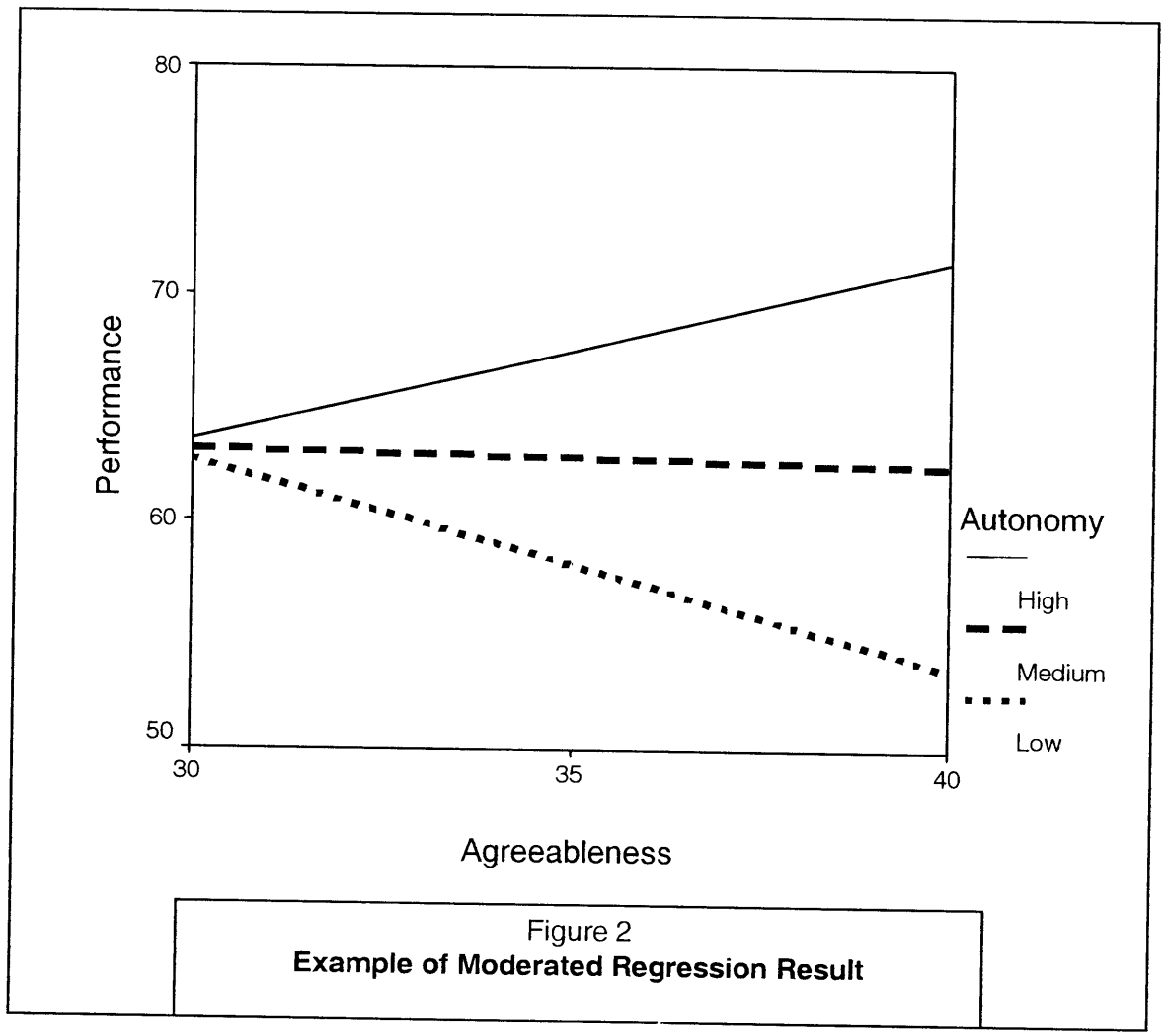

Figure 2
Example of Moderated Regression Result

Figure 2 illustrates moderated regression. Gelattly and Irving (2001) tested the hypothesis that job autonomy moderates the effects of personality on job performance. The hypothesis was supported for the Agreeableness scale of the NEO-FFI (Costa & McCrae, 1992). Figure 2 illustrates the resulting interaction by plotting the regression of performance on agreeableness for high, medium, and low autonomy. An interaction is indicated by the fact that the lines are not parallel.

The lines would be parallel if performance were simply an additive function of agreeableness and autonomy. The nonparallel lines mean that the relationship between personality and performance depends on the level of autonomy. This contingency indicates an interaction.

The moderated regression approach was not always accepted as a legitimate method of modeling interactions. A debate on the use of this method once focused heavily on how to provide appropriate statistical significance tests. Of central concern was the confounding of main effects with interaction effects (i.e., collinearity). Cohen (1978) facilitated widespread use of the method by showing that analyses produced the same conclusion regarding the presence of an interaction whether or not special steps (e.g., standardizing the variables) were taken to reduce this problem. In this special case, a crucial element of the principled debate could be resolved in purely mathematical terms. Such definitive argument resolutions are rare in behavioral research.

SEM includes analogues to both the ANCOVA and moderated regression methods of modeling interactions. Subgroup models provide the SEM equivalent of the ANCOVA approach. Subgroup analysis divides samples into groups based on some characteristic(s), such as occupation or gender. Two versions of an SEM then are fitted to the data in each group. One version constrains the model parameters to be equal for all the groups in the analysis. This step is equivalent to conducting ANCOVA assuming that a single regression line applies to all groups. Other versions permit differences in parameter values across groups. If removing the equality constraints yields substantially improved predictive accuracy relative to the first model, model parameters cannot be considered equal in the groups. This result is equivalent to finding nonparallel regression lines in ANCOVA.

The SEM analogue of moderated regression creates cross-product terms by multiplying scores on indicator variables. This method of testing for interactions is common in some areas of research (e.g., industrial-organizational psychology). A debate has been in progress in the SEM literature since Kenny and Judd (1984) first introduced the basic method. Subsequent work by Ping (1995, 1996) and Joreskog and Yang (1996) simplified the implementation of the original method. Topics under discussion today include how many interaction indicators must be included in a model and how to differentiate curvilinearity from an interaction. Significance tests in this area are sensitive to the underlying distributions of the variables. Schumacker and Marcoulides (1998) summarized available methods and the ongoing debate about how to best specify and estimate interactions.

Interactions and curvilinear relationships almost certainly will be continuing concerns in theory development and modeling. The current state of the art does not provide definitive direction regarding the best method of addressing these issues. Steiger (2001) recommends reading Rigdon, Schumacker, and Wothke (1998) and Joreskog

(1998) (Schumacker & Marcoulides, 1998) as context for other recent works in this area. Investigators may also wish to consider issues related to methodological limitations of applied research that affect the results from moderated regression analyses (McClelland, 1997; McClelland & Judd, 1993; Russell & Bobko, 1992; Russell, Pinto, & Bobko, 1991). These limitations may be even more problematic than the purely statistical issues.

At present, the multiple-group approach is the preferred option for modeling interactions in SEMs. Applications of this approach can be viewed as a multivariate extension of moderator analysis. The literature on that procedure may provide useful qualitative guidelines to help avoid some potential problems (e.g., Zedeck, 1971).

Latent Growth Curve Analysis. Growth curves are a special area of interest in SEM. The analysis of change is a longstanding problem (Harris, 1963). However, pattern of change must be analyzed to address many interesting and important research questions about adaptation and development. For example, how do individual recruits adapt to the psychological stress of boot camp? How do the attitudes that affect reenlistment develop over time? Can early patterns of change be used to predict success and failure in military service? Such questions can be addressed today by modeling change as growth curves.

The SEM approach to quantifying change was stimulated by Rogosa and his colleagues (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1985). Growth curve analysis can be applied when a variable of interest is measured at several points in time. Given multiple measurements, a growth curve expressing the level of the measured variable as a function of time can be fitted to the data for each individual. Fitting the function yields a set of parameter values for each individual. If growth were linear over time, the parameters would be the slope and intercept of a regression equation that applied to a specific individual. Those parameter values then can be treated as dependent variables in analyses that relate them to attributes of the individual, group membership, and other potential predictors. The growth curves fitted to data usually are fairly simple (e.g., linear growth), but many different curves are possible in principle (Rogosa & Willett, 1985).

SEM is not the only method of analyzing latent growth curves. Raudenbush and Bryk (2002; Raudenbush, 2001) have extended the growth curve approach to include the construction of multilevel models (cf. Hierarchical Linear Models [HLM]) discussed in the following section.

Hierarchical Linear Models

Factors affecting behavior often have a hierarchical structure. Consider, for example, the problem of modeling morale during military basic training. Morale could be measured at weekly intervals during training. Other variables would be measured to understand the factors

that affect morale. These variables could include recruit characteristics (e.g., age, gender, personality), unit membership (e.g., platoon, division, flight), and unit characteristics (e.g., number of recruits, average intelligence test scores, average experience of instructors). Models describing morale can be constructed from the combination of trends over time and a combination of individual and group characteristics.

HLM provides a general method of addressing research problems such as that described in the previous paragraph (Raudenbush, 2001; Raudenbush & Bryk, 2002). The basic problem is hierarchical because it involves several distinct levels of analysis. The first level consists of within-person processes. In the morale example, the within-person model would describe changes in morale over time for a single individual. Suppose morale was low early in training and then increased at a constant rate over time. This temporal pattern could be modeled by representing morale as a linear function of time. The model for each person would consist of the slope and intercept for the linear model. The linear model would be the latent growth curve for morale for that individual.

The second level would introduce individual differences as a factor in morale during training. Recruit characteristics could be used to predict differences in the slopes and intercepts estimated in the within-person level of analysis. The modeling process could be extended to a third level that characterized unit differences in morale. This level of the model would test for average differences in the intercept and slope of the within-person model. This level also would relate the average differences to the unit characteristics that had been measured. The HLM approach, therefore, makes it possible to analyze changes in morale as a combination of within-subject, between-subject, and between-unit effects.

HLM addresses two important limitations of traditional behavioral models. Traditional models commonly follow disciplinary boundaries that treat growth processes, individual differences, group dynamics, and social environment as distinct research topics. These distinctions can lead to incomplete or biased models when the behavior of interest is affected by factors at more than one level of analysis. In such cases, the predictive accuracy of single-level models is limited by the fact that some systematic sources of variance are omitted. The model is incomplete and therefore cannot fully account for the data. Focusing on a single level of analysis also can bias estimated effects for the current level of analysis. Bias will occur when variables in the single-level model are confounded with causal factors at other levels of analysis (cf., James et al., 1982). HLM provides tools to produce models with greater explanatory power by combining the several levels of causal factors. The same tools produce more accurate estimates of the effects of factors at each level controlling for the effects of factors at other levels.

The advantages of HLM can be critical to a proper understanding of behavioral phenomena. For example, Duncan, Jones, and Moon (1993) analyzed the health behavior of people living in different regions of the United Kingdom. Previous research suggested that health behavior differed between geographical regions. Those differences could be interpreted as evidence that regional cultural differences affected behavior. Duncan et al. (1993) used HLM to show that the reported geographical differences were almost completely explained by demographic differences between regions. In a very different context, Sliwinski and Hall (1998) applied HLM to assess claims that aging exerts a general negative effect on all mental capacities. Sliwinski and Hall's (1998) hierarchical model grouped the mental tests into categories. When those categories were included in the model, age effects were limited to just a subset of the mental capacities. These examples (and others, see Raudenbush & Bryk, 2002) illustrate that HLM can be applied to problems ranging from processes occurring within individuals to processes that characterize broad socioeconomic groupings.

HLM could be extremely valuable in military behavioral research. Previous applications of this methodology in educational research (cf., Raudenbush & Bryk, 2002) have obvious parallels to military research on education, but the potential extends well beyond this area of study. The key elements of the method are the availability of a series of measurements on individuals combined with a hierarchical structure of some type. The approach could be applied to topics such as evaluating different weight loss programs by determining growth curve parameters for participants. The relationship between average growth curve parameters and quantitative or qualitative characteristics of the programs could be analyzed to identify the critical ingredients of effective programs. These analyses could include adjustments for the effects of individual differences on growth curve parameters. The adjustment would make it possible to estimate program effects while controlling for differences in personnel composition.

The use of HLM could stimulate theory development by integrating models from traditionally distinct research disciplines. Theoretical statements from individual differences in psychology, social psychology, sociology, and organizational psychology can be combined. This end is achieved by developing models that treat the individual, the small group, and social categories as different levels in a hierarchy. Parameters for the various perspectives can be estimated in a single analysis. The estimation procedures adjust for differences at other levels within the model, so allowance is made for group composition when analyzing unit effects and vice versa (Raudenbush & Bryk, 2002).

The merging of theoretical perspectives could increase the importance of models of group dynamics in some key areas. For example, delinquent behavior (e.g., attrition, nonjudicial punishment) is more prevalent in some military units than others. If the

differences remain after controlling for unit composition, attention to explanations based on unit factors is a reasonable step. Military tradition makes it likely that leadership would become one focus of inquiry. However, leadership may not be the source of the unit differences (Vickers, Hervig, Wallick, & Conway, 1984). If so, alternative models must be considered. Behavioral contagion (transmission of negative attitudes from one person to another) is an example of a possible alternative (Jones, 1998). Properly applied, HLM can test individual difference, unit leadership, and behavioral contagion models at the same time. This joint representation of multiple levels of analysis in a single model may be needed to represent the complexity of the influences on real behavior.

Using HLM to avoid incomplete and biased models will introduce new research design issues. In general, the organizational units studied are only a sample from a population of similar units. This observation raises sampling issues. How many units should be sampled to accurately estimate effects? A large sample of individuals no longer guarantees adequate sampling. The sampling frame for a study must include units as entities to be sampled. If it is convenient, all of the unit personnel may be included in a study. However, covering an adequate sample of units could require collecting data from very large numbers of individuals. If the cost of collecting data from an individual is high, it may be necessary to use stratified sampling within units.

One aspect of HLM is noteworthy to avoid confusion when considering this method in conjunction with other methods discussed here. In the context of HLM, a latent variable is any unmeasured variable (Raudenbush & Bryk, 2002). Thus, references to latent variable modeling are not equivalent to references to latent trait modeling in SEM. Instead, latent variable discussions in HLM are more likely to address problems such as the imputation of missing data. In this case, the missing data comprise a latent variable in the sense that each case presumably has a specific value, but that value is not measured in the study. These superficially different uses of the term "latent variable" are not contradictory. Instead, the different applications of the term represent different instances of unmeasured variables. The development of an integrated perspective on latent variables is an ongoing topic of discussion in data analysis generally (Bollen, 2002).

Categorical and Limited Dependent Variable (CLDV) Models

A wide variety of models can be grouped together under the heading of CLDV models. These models deal with dependent variables that are categorical (e.g., pass-fail) or time-limited in some way (e.g., the observation period is stopped or participants die, move away, or refuse to continue participation). In each case, the nature of the dependent variable creates difficulties because it either is not linearly related to predictors or because error variance is heteroscedastic, or both.

Military researchers often encounter CLDV. Examples of dichotomous categories include pass-fail measures of success in training programs, attrition, and reenlistment. More complex categorical criteria are represented by specific reasons for attrition (e.g., medical, behavioral, administrative) or rank/rate at the end of a period of enlistment. Time-limited variables can occur when an individual is transferred from one unit to another or returned to civilian status prior to completing an enlistment. Accurate modeling of CLDV obviously is important.

Recent developments have produced an integrated approach to CLDV models. The approach has two fundamental elements (McCullagh & Nelder, 1989). Each model has a transformation component known as a link function. The link function generates transformed variables that have linear relationships to predictor variables. The choice of a link function depends on the nature of the dependent variable. The critical point is that the transformed variables can be analyzed using familiar linear regression methods.

The second element of CLDV models is known as the systematic component. This component is a linear model quantifying the relationships between predictors and the transformed dependent variable generated by the link function. CLDV models can be fitted using the general linear model (GLM) approach to obtain appropriate statistics.

The link function approach to CLDV has several positive characteristics (Long, 1997). The most important is that it is no longer necessary to settle for linear approximations to more complex relationships. Complex mathematical functions can be approximated by linear functions over narrow ranges of scores. In the context of CLDV, this fact means that a traditional linear model based on raw data can have acceptable predictive power even though it does describe the true functional relationship between the predictors and the dependent variable. The apparent predictive power can be misleading. For example, basic assumptions such as homoscedasticity of error variance can be violated. More importantly the resulting equation may give the impression that outcomes (e.g., risk of attrition) increase equally with each point in the scale score. In fact, the change may be much greater in some areas of the score range than in others. The appropriate mathematical expression must be employed to accurately assess effects and accurately predict outcomes.

Capitalizing on the advances in CLDV analysis also provides other payoffs. Link functions make it possible to apply familiar methods from regression analysis to build more robust models. For example, outlier and influential data points can be identified in CLDV analyses using the same procedures employed in linear regression. As another example, polynomial regression can be used to express the transformed variable as a nonlinear function of a predictor. The link function approach also makes it possible to apply maximum likelihood methods

for estimating parameter values. Thus, familiarity with this general approach to estimation helps construct CLDV models.

CLDV models cannot be interpreted directly. Link functions involve nonlinear transformations. The transformation must be reversed to predict the original variable. The reverse transformation must be considered because a change of one unit in a predictor can have very different effects in different score ranges. For example, in a logistic regression model, the effect of a one-point difference in scores is not the same over the full range of scores. At the high and low ends of the range, a one-point difference typically will have little effect on the predicted outcome. The same difference can have a pronounced effect in the middle range of the score distribution. This point should not be a barrier to the use of CLDV methods. The benefits of the approach make it worthwhile to take the time to understand its procedures.

One ongoing line of development in CLDV analysis is of special interest for model construction. The use of specific contrasts within a GLM approach to CLDV has been explored as a method of formulating and testing alternative causal paths (von Eye & Brandstadter, 1998). These causal paths include a "wedge" by which two causes can cause a single outcome, a "fork" by which a single cause can lead to different outcomes, and a "chain" by which a cause and an outcome are linked by an intermediate variable. This approach holds promise of making it possible to use procedures such as loglinear modeling to test very specific causal hypotheses expressed entirely in terms of categorical variables. The basic method draws on the construction of contrasts that are similar to those found in ANOVA, but the structure of the contrasts is linked to specific hypotheses about sources of causal effects.

Path Models and MAGIC

The methods described above can increase the quality of evidence available to principled argument process. The relationships between these methods and MAGIC criteria are too complex to describe in detail. However, some examples of potential gains can be provided.

Good explanatory power (*magnitude*) is a fundamental goal of modeling. This goal can be promoted several ways.

- The HLM discussion illustrated the potential for combining sources of variance that might be explored independently in traditional models. The combination should increase overall explanatory power. The combination strategy also reduces the risk of biased parameter estimates.
- SEM can provide greater insight into the true strength of associations. This gain is expected because SEM estimates are corrected for measurement error. The modeling implication is that a given model will not be preferred to an alternative

simply because the first is represented by more reliable measurements.

- Advances in CLDV analysis methods make it possible to replace linear approximations involving raw variables with appropriate functional relationships. For example, probit or logit analyses can be used in place of linear regression. These tools, therefore, have the potential to increase the magnitude of associations.

Articulation will improve. The task of accurately translating verbal statements of theory into mathematical equations is difficult (Blalock, 1969). The difficulty increases if the analysis cannot incorporate key elements of the theory. Key elements may include nonlinear or curvilinear relationships, equality or proportionality constraints on specific parameters, and interactions. The methods described here make it easier to build these details into models.

Generality is an empirical question, as noted earlier. The basic issue is whether a single model is appropriate for all entities (e.g., people or groups) or if specific models are needed for different classes of entities (e.g., men and women). The methods described here can be applied to test for generality. SEM methods of testing for generality have been alluded to in connection with measurement models. HLM provides a great deal of flexibility in specifying alternative models. This flexibility could be employed to contrast models that specify a single general equation with models that specify equations nested within different subgroups. Because HLM methods include the link transformations necessary to apply the CLDV approach (Raudenbush & Bryk, 2002), HLM tests for generalizability extend to those analyses as well.

Interest value may also be improved by the methods described here. The process of articulating the rationale for a model is likely to stimulate thinking about alternative models. Comparing plausible alternatives is more interesting than simply evaluating a given model in isolation. Given similar models, the use of methods that make it possible to isolate key differences between models will increase interest value. The methods described in this section can promote analyses that focus on critical parameters that differ between models. The effects of modifying those specific parameters may be modest in terms of the absolute fit of the model, but still could be important for choosing between alternative models. The methods described here also should increase interest value by creating stronger links between theory and the mathematical representation of the theory. The empirical representation of this linkage may be a link function that transforms a dependent variable.

Finally, the *credibility* of theoretical assertions should be improved by recent advances in statistical modeling. Clear articulation of links between theory and statistical models imply more convincing use of data to evaluate theories. Credibility is supported

by direct comparison of models based on different theoretical positions and by stronger empirical tests of claims for the generality of different models. Because this process reduces the likelihood of confirmation bias, credibility should increase even if the evidence does not clearly support one model over competing alternatives.

Model Evaluation

Once constructed, a model must be evaluated for its adequacy. Rigorous evaluation is critical to research progress. Evaluation includes an appraisal of the existing model and subsequent amendments to improve the model. This section of this chapter examines the state of the art in these two areas of behavioral research.

Two points must be kept in mind when considering the problem of model evaluation. First, statistical models are sets of parameter estimates. The coefficients in a linear regression are an example of a set of parameters. The parameters in a statistical model must be interpreted to obtain a substantive model. Indeed, interpretation is the main business of research (Kirk, 1996). Statistical models generally are evaluated in terms of statistical criteria. While this practice is reasonable, one should not lose track of the fact that model interpretation is a separate issue. This point raises two important issues concerning the use of statistical guidelines in model evaluation and selection. First, when considering alternative statistical models, *the best model by statistical criteria may not always be the most plausible.* Consideration should be given to rejecting this model if another model fits the data (nearly) as well and is more plausible substantively. Second, *model evaluation should include steps to explicate the relationship between statistical parameters and behavior.* This relationship may not always be straightforward. Taken together, these issues are a reminder that statistics are only one basis for choosing between alternative models.

Another important point is that behavioral phenomena are complex. Complete behavioral models would have to include many parameters to reflect this complexity. In practice, parameters that contribute little to the overall accuracy of the model are omitted. As a result, models are only approximations. This point is acknowledged explicitly in the SEM concept of errors of approximation (Browne & Cudeck, 1993; MacCallum, 2003). These small systematic errors mean that *behavioral models cannot be expected to predict behavior with perfect accuracy.* Instead, the modeling goal should be an acceptable approximation to a complete model. The stopping point for model development is reached when only small effects are omitted. Serlin and Lapsley (1985) refer to this endpoint as the "good enough" principle for model construction.

Model Appraisal

Two general criteria—statistical significance and explanatory power—serve as central points of departure for this section. The

central theme here is that it is important to properly interpret these familiar statistical concepts. The null hypothesis significance test (NHST) and common ES criteria are examined to illustrate the state of the art in traditional model appraisal practices. SEM appraisal methods are examined to illustrate an alternative approach that is likely to have an increasing influence in the future.

Significance Tests

Significance tests are the most common tool for model appraisal in behavioral research. Significance tests can be diffuse (omnibus) or focused. Diffuse significance tests involve multiple degrees of freedom (*dfs*); focused tests involve a single degree of freedom (Rosenthal & Rosnow, 1984). For example, tests for main effects and interactions in ANOVA often involve $df > 1$. Planned or post hoc contrasts decompose the omnibus test into a set of component ($df = 1$) tests.

Significance test procedures can involve an NHST or a strong significance test (SST). NHST procedures compare the sample estimate of a parameter value to a hypothetical value of zero. SST computations compare sample estimates to values derived from theory or prior research. SST will include zero values when appropriate, but such cases are expected to be rare (Meehl, 1978; Schmidt, 1996). Thus, the hypothesized parameter value(s) in SST ordinarily will be nonzero values.

The focused-diffuse distinction is less important for the present discussion than NHST-SST distinction. The major issue associated with the difference between focused and diffuse tests involves the number of significance tests performed. Diffuse tests apply a single test to evaluate a set of parameters. Focused requires multiple significance tests (i.e., one for each degree of freedom). The process of performing multiple significance tests increases the likelihood that at least one result will be significant by chance. Special procedures to deal with this problem can be used to control the inflated risk of mistaking chance for a true effect (Keselman, Cribbie, & Holland, 1999; Seaman, Levin, & Serlin, 1991). Various methods for planned and post hoc comparison in ANOVA are examples of how to deal with this problem (Winer, Brown, & Michels, 1991). The following discussion assumes that appropriate procedures will be used to control for multiple significance tests.

NHST and SST have very different implications for modeling. Roughly speaking, NHST evaluates whether an attempt to construct a model is worthwhile. SST determines whether an existing model (i.e., set of parameters) should be retained. NHST and SST procedures are complementary when viewed as elements of an overall model development process. Initially, NHST is employed with the null hypothesis as a straw man model (Krantz, 1999). If the null hypothesis cannot be rejected, zero is a reasonable estimate for the parameter(s) being evaluated. A model in which all parameters are zero will be of

substantive interest only insofar as it rules out some possible relationships. Thus, NHST is most likely to be performed in connection with attempts to develop initial estimates of parameters that are believed to have nonzero values. As further research is conducted and evidence accumulates, refined estimates of model parameters can be developed based on the cumulative research evidence. SST can be applied when the model is represented by a set of nonzero parameter estimate(s) based on prior work or theory. In this situation, the question is whether the sample estimates of the parameters are close to the predicted values. Using the traditional $p < .05$ statistical significance criterion, the desirable NHST outcome is $p < .05$. This outcome justifies tentatively adding the parameter(s) being evaluated to the behavioral model. The desirable SST outcome is $p > .05$. This outcome would justify retaining the existing model.

NHST Procedures. NHST is the most common metric for model appraisal (Finch, Cumming, & Thomason, 2001; Kirk, 1996; Vacha-Haase & Ness, 1999). Meehl (1978) has argued that reliance on NHST is one reason for the stunted growth of behavioral models. His viewpoint is a widely quoted anchor point in an ongoing debate. Arguments in the debate range from recommending that NHST be banned entirely to arguing that NHST would have to be invented if it did not already exist (Abelson, 1997). The *American Psychologist* recently published a negative view (Cohen, 1994), followed by a rebuttal (Hagen, 1997), and an attempt at synthesis (Kreuger, 2001). The scope of the debate is broadened by examining topics such as the actual use of NHST in practice (Nelson, Rosenthal, & Rosnow, 1986) and the historical development of NHST (Cowles & Davis, 1982; Smith, Best, Cylke, & Stubbs, 2000). The full range of topics considered in the NHST debate can be found in Harlow, Mulaik, and Steiger (1997). A comparison between this and an earlier collection by Morrison and Henkel (1970) provides insight into the rate at which the debate has progressed. Nickerson (2000) also provides a brief comprehensive summary of the current status of the debate.

NHST can be a trap for the unwary. Cohen (1994, p. 997) highlighted this problem when he wrote: "What's wrong with NHST? Well, among many other things, *it does not tell us what we want to know* …" (italics added). Researchers collect data for the purpose of testing models. NHST results can lead to erroneous inferences about the status of a model for any of the following reasons:

- The NHST p value is not the probability that the model is correct. Instead, p is the probability of the data if the null hypothesis is correct. The critical point here is that the p value must be combined with other information to determine how the data relate to the probability of the model. Cortina and Dunlap (1997), Dixon and O'Reilly (1999), Krueger (2001), and Trafimow (2000) discuss Bayes's theorem as the appropriate method for using p as one element in estimating the probability that the model is correct. Howard, Maxwell, and

Fleming (2000) compare the Bayesian and NHST approaches. For the present purposes, it is sufficient to note that the relationship is not straightforward. For example, the null hypothesis can be rejected when the data actually increase the probability that this hypothesis is true (Lindley, 1957).

- The complement of the NHST *p* value (i.e., 1 - *p*) derived from a single study is not the likelihood that the alternative model is correct. The complement is not the likelihood that the results will replicate. Both interpretations are wrong, although NHST *p* values can be a rough guide to the likelihood of replication (Greenwald, Gonzalez, Harris, & Guthrie, 1996).

- Rejecting the null hypothesis in each of several studies does not mean their results were replicated. If the sign of the statistic used in the test was the same in each study, the results replicate qualitatively. This qualitative criterion is accepted as evidence of replication under NHST (Greenwald et al., 1996). However, a quantitative replication criterion could produce a different conclusion. For example, suppose three studies were conducted with $N = 200$ in each study. Suppose the correlations in the studies were $r = .15$, $r = .50$, and $r = .90$. The null hypothesis would be rejected in each study. However, most researchers would be reluctant to treat the results as equivalent because every pairwise difference would be statistically significant.

- NHST does not indicate whether a particular parameter is large enough to be important in practical or theoretical terms. The conceptual definition "Significance = Effect Size * Sample Size" (Rosenthal & Rosnow, 1984) shows why. Even trivial deviations from zero will be statistically significant given a large enough sample. Conversely, effects that are large enough to have practical and/or theoretical value will be statistically nonsignificant if the sample is small enough.

These interpretive pitfalls can be avoided by careful use of NHST. Harlow (1997) provides a succinct summary of options that are available to minimize the risk of misinterpretation. However, it is not easy to maintain perfection in this regard. Cohen (1994) lists an impressive array of established statistical experts who have erred at one time or another.

The list of things that NHST does not tell us is impressive, so why take the risk? The answer lies in the fact that NHST really is necessary in some instances. NHST is appropriate for evaluating whether findings are due to chance (Mulaik, Raju, & Harshman, 1997). NHST also is informative in answering some specific questions that involve dichotomous alternatives (Abelson, 1997; Greenwald et al., 1996; Hagen, 1997; Mulaik et al., 1997; Wainer, 1999). These applications of NHST support the argument that this procedure is a

necessary if sometimes misleading tool for model evaluation (Abelson, 1997).

The recommended strategy for minimizing the negative effects of NHST is to report results more completely (Meehl, 1997). A confidence interval (CI) is the most common recommended alternative to NHST for this purpose. This interval provides a point estimate of ES and indicates the precision of the estimate (Cumming & Finch, 2001; Greenwald et al., 1996; Wilkinson & the Task Force on Statistical Inference, 1999). CIs are directly linked to the familiar NHST procedures and support the development of cumulative parameter estimates as a research domain matures (Cumming & Finch, 2001). Methods of computing confidence intervals are available for all common ES indicators (Algina & Moulder, 2001; Cumming & Finch, 2001; Fan & Thompson, 2001; Fidler & Thompson, 2001; Mendoza & Stafford, 2001; Smithson, 2001). At a minimum, investigators should report the exact test statistic or exact significance level along with sample size (e.g., $t = 2.88$, 32 *df*). This information generally is sufficient to permit computations of ES and CI. The ES component of the CI leads the discussion directly to the second criterion for evaluating models.

SST Procedures. SST avoids some NHST problems by replacing the NHST assumption that ES = 0 with ES = k, where k is a parameter value that differs from zero. While k could be based on theory, behavioral theories seldom are sufficiently developed to permit this. Parameter values are more likely to be derived from prior research. SST, therefore, can be viewed as a consistency test. Are the current data consistent with the evidence from prior studies? If $p > .05$, this question can be answered affirmatively. If the sample were large, the range of parameter values that would yield an affirmative answer would be small. If the model is not correct, observed values that were close enough to the predicted values to fall in the range of acceptable values would be "a darned strange coincidence" (Salmon, 1984). As a result, the SST would be a risky test of consistency between the present data and either prior research or theory because most parameter values would be inconsistent with the model prediction (Meehl, 1990a). A coincidence that is consistent with a risky prediction provides strong support for the model being tested.

SST and NHST are formally similar. Both tests estimate the probability that the study results would have been obtained under a particular model. NHST asserts that the parameters in the model are equal to zero. SST specifies non-zero values. This difference is the reason that NHST and SST are complementary in the context of overall research programs. SST cannot be used without knowledge of the parameter values, so this procedure is not feasible in the initial stages of the study of behavioral phenomena. SST can be used once research provides non-zero values for the parameter estimates. At this point, NHST would be counterproductive because it ignores prior findings. Thus, replacing NHST with SST implies movement along the continuum from exploratory to confirmatory models. Movement toward SST is desirable because it implies stronger theory based on

cumulative empirical evidence. Movement toward SST should facilitate the development of reliable knowledge. Meta-analysis provides methods of accumulating results across studies (Glass, 1976; Glass, McGaw, & Smith, 1981). This analytic methodology is widely used at present, but meta-analytic results do not appear to be used to generate SST with any frequency.

In the final analysis, neither NHST nor SST is an entirely satisfactory method for model evaluation. Neither procedure addresses the fundamental question of whether the model is sufficiently accurate to satisfy Serlin and Lapsley's (1985) "good enough" principle. NHST is not satisfactory because the null hypothesis can be rejected when a model has virtually no explanatory power provided the sample is large enough. Similarly, the existing model associated with SST can be accepted even though it meets the criteria for a risky test. This can happen even with a large sample if the model parameters are known with some accuracy but represent only a subset of the parameters required for a complete model. The accuracy of the model is a distinct issue that can only be addressed by considering an additional criterion, explanatory power.

Explanatory Power

Explanatory power is how well the model accounts for variation in the phenomena of interest. This model attribute often is evaluated in terms of proportional reduction in error (PRE). PRE reflects the proportional reduction in cumulative error achieved by substituting the predictions from a fitted model for the predictions from the null model. Common PRE indices are r^2 for correlation, R^2 for regression, and ε^2 for ANOVA. Draper and Smith (1998) and Cohen, Cohen, West, and Aiken (2003) provide excellent introductions to explanatory power in relation to applied regression procedures. Their sections on model fit and related topics should apply to various types of GLM models. For example, computer programs often print out ANOVA tables for regression models and estimates of R^2. PRE measures also are available for models with categorical dependent variables (Hildebrand, Laing, & Rosenthal, 1977).

Explanatory power is linked to ES. The linkage makes it possible to express explanatory power in terms of either strength of association (e.g., R^2, ε^2) or magnitude of ES (e.g., *r*, Cohen's *d*). Both association and magnitude indices are readily available for common analysis procedures (e.g., regression, ANOVA, cf., Cohen, 1988; Hedges & Olkin, 1985; Kirk, 1996).When reporting ES or PRE, several points should be kept in mind:

- *Dichotomous decision rules are counterproductive.* The limitations of this approach are evident from the history of NHST. NHST procedures were developed in the context of the need to choose between alternative courses of action (Cowles & Davis, 1982). Significance standards were rule-of-thumb criteria established by

well-informed individuals who recognized a need to make a yes-no decision in the presence of uncertainty. The extensive literature on NHST demonstrates the problems that arose when this procedure subsequently was codified and ritualized (e.g., Meehl, 1978, 1990b). Flexible reasoning will be more productive than rigid application of a dichotomous decision-making scheme. Thus, Cohen's (1988) ES guidelines should be applied in the spirit in which they were offered. Transforming these guidelines into rigid rules for dichotomous decisions would be a serious mistake.

- *Small ESs can be important.* In applied research, small effects can be important when they involve repetitive events that yield large cumulative trends (Abelson, 1985) or when the outcome being predicted is very important (e.g., heart attack mortality; Rosnow & Rosenthal, 1989). In theoretical studies, small ESs can be important when there is a small difference between stimuli that produce an effect and/or when the dependent variable is difficult to influence (Prentice & Miller, 1992).

- *Capitalization on chance inflates sample estimates of explanatory power.* When parameters are estimated using data from a single sample, the analysis procedures are designed to maximize the fit of the model to the data. The maximization process capitalizes on chance elements of the data. As a result, the model will not fit the data from a new sample as well as it did the data from the original sample. The loss of predictive power is known as shrinkage. Methods of adjusting for shrinkage have been developed to obtain more realistic estimates of the predictive power that can be expected when a model is applied to a new data set. For example, the shrunken R^2 for regression and the ω^2, a comparable statistic for ANOVA (Hays, 1963), allow for this inflation. Joreskog and Sorbom's (1981) adjusted GFI is an SEM analogue of the shrunken R^2. Raju, Bilgic, Edward, and Fleer (1997, 1999) reviewed and simulated the performance of a number of equations for shrunken R^2. In their simulation, shrinkage increased as the predictive power of the model decreased, as the sample size decreased, and/or as the number of predictors in the model increased. These model components had more effect on shrinkage than did the choice between alternative shrinkage equations. These findings should generalize to other GLM analyses (e.g., ANOVA models). Thus, investigators should be especially concerned about shrinkage when a model with many predictors yields moderate to low predictive power in a small sample. Browne (2000) provided a general discussion of shrinkage and the available methods of adjusting for this capitalization on chance.

- *The choice of ES should be appropriate to the modeling objective.* For example, in regression, the semipartial correlation expresses PRE relative to the overall variance in the dependent variable. Significance tests are based on the partial correlation, a statistic that relates incremental PRE to the residual variance

(cf., Cohen & Cohen, 1983, pp. 85-110). When the overall model accounts for a large proportion of the criterion variance, the semipartial correlation can be small even though the partial correlation is large. The partial correlation is the basis for NHST. This statistic relates the incremental variance accounted for by adding the parameter to the residual variance for the overall model. The semipartial correlation expresses the incremental variance relative to the overall variance of the dependent variable. For example, if a model accounted for 90% of the variance in a dependent variable, a parameter that accounted for 10% of the residual variance would only account for 1% of the total variance. The model is being constructed to explain the overall variance, not the residual variance. The semipartial correlation indicates the explanatory power of the model in this context, and so would be more appropriate than the partial correlation for most modeling situations.

Interpretation is a problem for model appraisals based on traditional indices for explanatory power. Problems arise because ES and PRE indices are set in a statistical frame of reference. In each case, raw data are transformed into standardized data. The advantage of transforming the data is that ES values can be compared even when different variables in the model have different raw score metrics. For example, analysis might yield an ES represented by a point biserial correlation between experimental status (i.e., experimental or control group) of $r_{pb} = .30$. The associated PRE statistic would describe the relationship as accounting for 9% of the variance in the dependent variable. Cohen's (1988) criteria would classify the association as moderate in size. These statements could be applied whether the experiment was a training program designed to increase push-up scores, a clinical intervention to reduce depression, or a new method of teaching designed to improve algebra test scores.

The disadvantage of ES-based model appraisal derives from the transformation of the raw data. The standardization must be reversed to express ES in behavioral units relevant to the original research question. Commentaries that contrast statistical significance with practical or theoretical significance highlight this necessity (e.g., Jacobson, Roberts, Berns, & McGlinchey, 1999; Thompson, 2002). Solutions include the binomial effect size display (Rosenthal & Rubin, 1979), the common language ES (CL; McGraw & Wong, 1992), the receiver operating characteristic curve (Lett, Hanley, & Smith, 1995; Swets, 1988), and the number needed to treat (Ebrahim, 2003). For example, CL is the probability that an observation selected randomly from an experimental group will perform better than an observation selected randomly from a control group. Thus, CL = 75% means that a comparison between the two observations will favor the experimental group 75% of the time. This result has clear intuitive meaning. Also, the difference between CL = 75% and CL = 53% is immediately meaningful. The other indices mentioned here provide comparable translations of ES into the behavior(s) of interest. The use of these indices should

allow for uncertainty in the ES estimates. This allowance could take the form of CIs expressed in a CL ES metric.

Consistent reporting of ES would support the growth of reliable knowledge in behavioral research. Improvement in this aspect of statistical practice would ensure that enough information was reported to support meta-analysis of the cumulative body of evidence in a field. Meta-analysis can formally model methodological and substantive influences on ES. Several meta-analytic methods developed for this purpose (Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1984) produce similar results (Schmidt & Hunter, 1998).

Meta-analysis generally is used to evaluate correlations or other ES measures that can be converted to correlations (Cooper & Hedges, 1994; Hedges & Olkin, 1985). However, meta-analytic methods can be extended to parameters such as standard deviations and standard errors of estimate (SEEs) (Raudenbush & Bryk, 2002, chapter 7). These extensions should receive increased attention in future meta-analyses. The variables that predict ES in a meta-analysis are analogous to moderators in traditional moderator analysis. Restriction of range and other factors can produce the appearance that a moderator effect is present when it really is not (Zedeck, 1971). Meta-analysis, too, can be influenced by these factors. Extending meta-analysis to cover sampling variance reduces the risk of incorrect inferences. With this extension, meta-analysis can provide parameter estimates that are suitable for SST. These estimates would move behavioral research toward risky hypothesis tests that could provide the evidence needed to make strong claims for a model.

Full realization of the potential value of meta-analysis may be hampered by the appearance that meta-analysis is too complex for the average researcher. This appearance is misleading because the basic analysis procedures are no different than those found in primary data analysis (Rosenthal & DiMatteo, 2001). Special issues that are unique to meta-analysis are described in Cooper & Hedges (1994). Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a framework for thinking about the combinations of methodological and substantive factors that may have to be combined to produce complete models to account for variations in ES across studies.

The Future of NHST and ES

The preceding comments identify opportunities to improve on current practices by reducing the emphasis on NHST and increasing the emphasis on ES and PRE indices of model effectiveness. Consistent reporting of CI would support movement toward SST by facilitating meta-analytic summaries that would provide the parameter estimates needed for SST. That shift will provide the basis for stronger models based on cumulative evidence, rather than on a study-by-study analysis coupled with discussions that provide qualitative comparisons to prior

findings. The development of models based on formal analysis of the cumulative empirical evidence should foster consensus on the evidence.

Movement toward the use of CL ES metrics is another factor that should foster consensus. Developments in this area would promote a better understanding of what the set of parameters in a model mean in terms of actual behaviors that are the true focus of model building. The gap between abstract statistical indices and actual behavior is clear and must be addressed in practice. Consensus on the choice between alternative models will not result directly from these changes in practice, but it is reasonable to hope that the bases for arguments about different models will be communicated to practitioners more clearly.

ES will be reported more consistently in the future. Melton's (1962) editorial on significance tests is commonly cited as evidence of the pressures that made NHST an important, sometimes critical, requirement for publication. Similar pressures are mounting for ES indicators that are reported sporadically at present (Finch et al., 2001; Kirk, 1996; Vacha-Haase & Ness, 1999). A growing number of journals have editorial policies that require additional information (Fidler & Thompson, 2001). Previous experience suggests that change may be slow (cf., Finch et al., 2001), but the increasing frequency of meta-analyses should stimulate more consistent reporting. The work of researchers who do not report ES – or who fail to provide enough information to compute ES – will ultimately be excluded from the cumulative body of evidence.

Two trends should foster improved inference about the adequacy and utility of models. Interpretation may be improved by combining progress toward clinical ES measures with the recommended use of CI. These approaches could be combined to present findings graphically in units that have direct clinical or applied meaning and utility. Graphical presentation that fosters better communication of research findings is one index of the scientific maturation of a field (Smith, Best, Stubbs, Archibald, & Roberson-Nay, 2002). Both trends should decrease the need for practitioners to apply arbitrary statistical standards when making judgments about the behavioral implications of models.

The increased use of Bayesian statistics will also support improved inference. Elements of Bayesian reasoning already are present in some current analysis methods (e.g., HLM; cf. Raudenbush & Bryk, 2002). The frequency with which Bayesian reasoning is discussed in the NHST debate may increase familiarity with this approach to inference. The problem of how to specify prior probabilities is the primary barrier to wider use of Bayesian models. Recent summaries of the average ES from multiple meta-analyses (Lipsey & Wilson, 1993; Meyer et al., 2001) may provide some leverage for this problem. These summaries provide an empirically grounded a priori estimate of the prior distribution of ES for behavioral research. Stein's paradox (Efron & Morris, 1977) can be applied to invoke this distribution as a

proxy for the true priors in new research domains. Thinking along these lines may replace NHST with more appropriate inferential thinking.

One undercurrent in the NHST debate merits special mention to close this topic. Data analysis should not be a ritual. Data analysis is only one element in the overall process of empirically testing hypotheses and models. Existing theory and prior research findings should guide the process at all times. Judgment is needed at each step in the research process to produce designs and analyses that correctly link data to research questions. Researchers routinely use judgment in the complex activities of formulating hypotheses and developing research designs (Kirk, 1996). The best overall statement regarding NHST at present, therefore, appears to be this: *Judgment should not be suspended during the data analysis phase of research.*

SEM Appraisal Methods

Traditional model evaluation methods will persist until a reasonable alternative is available. SEM practices are considered in some detail here because they are the product of a quarter century of developing an approach to model evaluation that minimizes reliance on significance testing. Also, the increasing use of SEM in behavioral research demonstrates the attractiveness of these methods. Researchers should be motivated to learn new model appraisal techniques in the process of acquiring familiarity with this new analytic methodology.

General Appraisal Processes

SEM appraisals involve three general criteria. SEM analogues of significance tests and explanatory power are coupled with indicators of misfit between models and data. Significance tests play a minor role in SEM appraisals. In this context, the confounding of ES and sample size has been an explicit concern for 20 years (Hoelter, 1983). Recognition of this fact has limited the use of significance tests primarily to the assessment of individual parameters within models. Parameter evaluation typically employs Joreskog and Sorbom's (1981) $t \geq 2.00$ criterion. This criterion approximates the $p < .05$ standard commonly used in NHST. This practice is primarily important in deciding model details rather than in evaluating the model as a whole. Earlier comments on NHST and SST apply to this element of SEM appraisal and will not be repeated here.

SEM programs report more than 20 GFIs that describe the overall fit between the model and the data. Classification schemes based on conceptual and/or computational similarity have been developed (e.g., Arbuckle & Wothke, 1999; Tanaka, 1993). However, simulation studies have shown that different GFIs are correlated when compared across samples. The empirical pattern of associations suggests two general GFI categories (Hu & Bentler, 1998). One category contains SEM

analogues of PRE indices. Cross-validation indices fall within this category. The second GFI category consists of measures that are analogous to SEE in regression analyses. The empirical clustering of GFIs is one reason for current recommendations that investigators report more than one GFI when evaluating SEMs. The recommended practice is to report at least one index from the clusters analogous to PRE and SEE (e.g., Bentler & Dudgeon, 1996; McDonald & Ho, 2002).

There is not yet a strong consensus on the best PRE measure for SEM. The RMSEA (Arbuckle & Wothke, 1999) has been recommended (Bentler & Dudgeon, 1996; Hu & Bentler, 1998). RMSEA has a population interpretation, so CIs can be computed. The population interpretation of RMSEA can be used to test hypotheses about the fit of the model. SEM programs often report a "p(close)" test that compares the observed RMSEA with a null hypothesis of RMSEA = .05. Fabrigar et al.'s (1999) recent recommendation that RMSEA should be used in the evaluation of EFA may indicate movement toward a consensus.

The standardized root mean square (SRMR) is the recommended SEM analogue of SEE (Bentler & Dudgeon, 1996; Hu & Bentler, 1998; MacCallum & Austin, 2000). SRMR reflects the standardized difference between observed covariances and the model estimates of those covariances.

Simulation studies indicate that RMSEA and SRMR provide different types of information about models. In these simulations the experimenter defines the true population model. Models with known errors (i.e., omitted parameters, added parameters) then are fitted to the data. GFI measures are evaluated by determining how sensitive they are to the errors. In such simulations, SRMR has been sensitive to errors in the path model component of SEMs in simulations (Hu & Bentler, 1998). RMSEA has been sensitive to errors in the measurement model (Fan, Thompson, & Wang, 1999; Hu & Bentler, 1998). Neither the number of factors nor the number of indicator variables affects RMSEA when the model is correctly specified (Cheung & Rensvold, 2002).

Users should be aware that RMSEA and SRMR can yield different conclusions about a model. This is not surprising given that these indices provide different types of information. Browne, MacCallum, Kim, Andersen, and Glaser (2002) describe the conditions that produce this disparity. When conflicts occur, trade-offs between these criteria may be required. For example, the available simulation evidence might be used as a guide. If so, SRMR would be given greater weight when evaluating path models. RMSEA would be given greater weight when evaluating measurement models. This approach to weighting the criteria assumes that models must lead to the adoption of a single model. One alternative would be to treat the criteria as equivalent and conclude that the study did not make it possible to choose between the model with the smallest RMSEA and the model with the smallest SRMR. Retaining more than one model may be preferable to premature adoption of a single alternative as "the" model.

SEM Appraisal Issues

As mentioned at the beginning of this section, SEM appraisal practices raise issues that are not always evident in other types of analysis. As a consequence, SEM appraisal does not begin and end with the examination of one or two statistical indicators for model adequacy. Satisfactory assessment of a model must also consider other issues. Some important general topics in model evaluation are examined here under the heading of appraisal issues.

Steps in Modeling. One important model appraisal issue is highlighted by recommendations that measurement models be defined and evaluated before estimating path models (Anderson & Gerbing, 1988). The initial proposal of this two-step procedure stimulated debate on the strengths and weaknesses of the approach (Anderson & Gerbing, 1992; Fornell & Yi, 1992). McDonald and Ho (2002) raised the issue again and demonstrated that the fit of the two models can be quite different. That demonstration should spark renewed interest in the topic.

Good overall fit for a model means that it reproduces at least some parts of the data well. However, overall fit can conceal significant misfit in specific elements of the model when a few large errors are averaged with a number of much smaller errors. If the large errors are scattered throughout the covariance or correlation being analyzed, there may be no problem. However, there is no guarantee that the errors will not be focused in specific areas of the model. Inaccuracies in the measurement model do not have the same implications for theory as do inaccuracies in the path model. A weak measurement model means that the current model does not satisfy one of several conditions that must be met to obtain meaningful tests of substantive hypotheses (Meehl, 1990a). The hypothesized relationship still might be demonstrated by refining the measurements or by substituting other measurement procedures if available. In fact, demonstrating that the same associations and lawful relationships between theoretical constructs can be derived using different measurement models is one hallmark of reliable knowledge (Ziman, 1978). This point is not always appreciated in behavioral research. Katzko (2002) argues that research paradigms often come to be equated with the theoretical constructs they are intended to measure. A general construct thereby is reduced to a specific set of operational definitions, including a specific measurement model. When different researchers develop different paradigms to study the same construct, each paradigm can become the center of a research program. Different programs then proceed in parallel rather than being directly compared. The collected set of paradigms then may be combined to represent the theoretical construct as a syndrome. Separate measurement and path models would help to clarify the role of measurement paradigms: Do different paradigms produce equivalent estimates of the relationships between theoretical constructs? If so, progress is being made toward reliable knowledge. In this context, measurement methods are

auxiliary models that must be reasonably accurate in order to test theoretical assertions (Meehl, 1990a).

A two-step evaluation is implicit in current practices outside the SEM realm. Regression and ANOVA methods typically define predictor and criterion measures prior to analysis. Consideration of the measurement model may be limited to reference to previous studies that established the measurement adequacy of the scales. Direct demonstration of measurement adequacy in the present sample is not ordinarily undertaken. More consistent attention to this issue would reduce the risk of inappropriate generalization. This risk is a neglected dilemma for behavioral researchers (Blalock, 1982). Neglect renders the dilemma invisible, but does not eliminate it.

Considered in this context, any process of principled argument must focus on measurement issues at some point. Routine use of a two-step analysis procedure can organize the empirical evidence bearing on this part of the argument. For this reason, it seems likely that practice ultimately will move in the direction of separating measurement model assessments (i.e., scale construction) from path model assessments (i.e., tests of substantive hypotheses). A review of Anderson and Gerbing's (1988) arguments, subsequent debate (Anderson & Gerbing, 1992; Fornell & Yi, 1992), and McDonald and Ho's (2002) recent exposition of the issues will provide researchers with a firm basis for determining how critical this issue is to any particular research problem.

Effects of Measurement Error. SEM evaluations also direct attention to the effects of measurement error. SEM programs provide R^2 values for latent traits that are dependent variables in the path model. The R^2 values are likely to be stronger than those found in ordinary regression. This difference can be attributed to removing the effects of measurement error (Bollen, 1989). In effect, SEM includes corrections for the attenuation of associations that result from measurement error (cf., Nunnally & Bernstein, 1994). The resulting estimates may be closer to true population values than are the attenuated estimates, but this apparent benefit should be viewed with caution (Bedeian, Day, & Kelloway, 1997). This potential advantage of SEM analyses is not always evident because R^2 for the path model does not ordinarily receive as much attention as it would if the model had been created using regression techniques. These statistics ordinarily play little part in SEM model evaluations. For example, the change in R^2 resulting from dropping a parameter ordinarily is not a consideration. The potential value of greater attention to these statistics is uncertain because the R^2 indicators of strength of association between theoretical constructs often are not reported even when they should be (Boomsma, 2000; McDonald & Ho, 2002). More consistent attention to this information in the future will provide a better basis for assessing the utility of the path model R^2 as an SEM criterion.

Search for Areas of Misfit. The appraisal of model misfit should include a search for atypical data points. In some cases, a few unusual observations, known as outlier and/or influential data points, heavily influence model fit. Roughly speaking, an outlier data point is an observation with an exceptionally large residual. The exceptional residual inflates the cumulative error variance of the model. The inflation means the model's explanatory power/goodness of fit is underestimated. Influential data points markedly alter parameter estimates in the model. For example, an influential data point is one that changes the regression slopes in multiple regression. The influential data point is not necessarily an outlier because distorted parameter values make the predictions reasonably accurate in some cases. However, the model parameters are less accurate than they could be for other data points. The overall accuracy of the model is likely to decline. Both outlier and influential data points can lead to models that do not accurately describe the population under investigation. In the context of behavioral modeling, the resulting model may lead to mistaken conclusions about the causes and consequences of the behavior of interest.

Outlier and influential data points are not fatal problems in modeling. Diagnostic procedures are available to identify influential and outlier cases (cf., Belsley, Kuh, & Welsch, 1980; Stevens, 1984). Chatterjee and Yilmaz (1992) review the use of these methods and provide an additional example of their application. These methods are available in many regression programs, but are not generally available in SEM or HLM programs. In those cases, preliminary regression analyses can help to identify exceptional data points. The sources that describe the bases for the diagnostic indicators provide general guidelines for interpreting the statistics. One limitation of the available indicators is that they may be insensitive to situations in which groups of data points affect the model (Belsley et al., 1980). Robust regression produces accurate models even when the proportion of contaminating data points is large (Rousseeuw & Leroy, 1987). This method should be considered for data screening when it is available, but the technique is not widely available at present. Draper and Smith (1998) describe an iterative approach that addresses this problem without the need for specialized analysis packages. The prediction of sum of squares (PRESS) approach is potentially time-consuming because it is iterative, but the effort may well be worthwhile.

Outlier and influential data points may even have positive effects in modeling. These exceptional data points can indicate that the data include cases that represent two or more distinct populations (Barnett & Lewis, 1994). If so, separate models can be constructed for each population once appropriate indicators of group membership have been identified.

Justification for Model Amendments. The model defined at the outset of a research project seldom is wholly satisfactory. The

appraisal typically identifies weaknesses that cannot be attributed entirely to chance or to exceptional data points. Investigators then must choose between amending the model and accepting it as "good enough."

The "good enough" choice should not be neglected (Serlin & Lapsley, 1985). This option is more likely to be considered in SEM modeling than in other areas. SEM practices set stopping rules in terms of GFI criteria. The most common cutoff for acceptable fit is GFI ≥ .900 (Bentler & Bonett, 1980), but higher standards (i.e., GFI ≥ .950) have been suggested recently (Hu & Bentler, 1999). In either case, less that perfect fit is considered acceptable. In practice, the standards for accepting a model as adequate are somewhat lower. Optimistic interpretation of fit indices is common (Bentler & Dudgeon, 1996). Care must be taken to ensure that the criterion for "good enough" is not set so low that it impedes the search for improvements in mediocre existing models.

Post hoc model modifications take two forms. The most common is the addition of parameters to improve the predictive accuracy of the model. Additions are philosophically defensible (Meehl, 1990a), but the modification process must be sensitive to the risk of capitalizing on chance. For example, in SEM, a search through all the constrained parameters is likely to capitalize on chance (MacCallum, Roznowski, & Necowitz, 1992). The same problem arises in regression (Thompson, 1995). Decisions regarding model modifications should be sensitive to the effects of chance. The basic approach is to set a more extreme significance standard (e.g., $p < .01$ rather than $p < .05$). Methods of testing post hoc contrasts in ANOVA may be the most familiar example of this approach (Winer et al., 1991). Keselman et al. (1999) and Seaman et al. (1991) compare several approaches that could be used in regression. Green, Thompson, and Poirer (2001) demonstrate the utility of this approach in SEM.

Model amendment should not rely solely on significance tests. Modifications should also consider ES and theory. Kaplan (1990a, 1990b) proposed combining ES and significance (i.e., modification index) to determine when to add parameters to SEM. Experts who commented on this proposal noted that parameters should not be added without sound theoretical justification (Bollen, 1990; Steiger, 1990). The rationale for that assertion applies to all types of models.

Models can also be modified by deleting parameters. Deletion fixes a parameter that had a non-zero value in the original model at zero in the revised model. In SEM, parameters with $t < 2.00$ often are deleted (Joreskog & Sorbom, 1981). Parameter deletion procedures can be implemented easily in regression. Backward stepwise regression performs these modifications automatically by removing the predictor with the least predictive value and then estimating a new regression with the remaining variables. This process is repeated until all remaining parameters are acceptable (e.g., $p < .05$). Backward

stepwise regression provides an analogous model amendment procedure for regression models.

Deletions reduce the number of parameters in a model. The result is greater parsimony (i.e., a model with fewer parameters; cf. Popper, 1959). However, the search for simplicity should apply the same principles used when deciding whether to add parameters. The effect on predictive power and the implications for theory should be considered. The effects of chance also should be considered. When examining a set of parameters, one or more of the parameters can appear to be no greater than zero by chance. Allowances should be made for this risk just as one would allow for chance effects when adding a parameter. This problem does not appear to have been addressed in the literature, but it is likely that the methods used to avoid improper addition can be adapted to avoid incorrect deletions.

The preceding sketch of model modifications points to two general principles. First, multiple criteria should be used when deciding whether to amend a model. The criteria include statistical significance, ES, and theory. Theory must be emphasized to avoid letting the statistical tail wag the theoretical dog. Second, the same criteria apply to additions and deletions from the model. However, significance tests should focus on Type II error (vs. Type I error) when considering deletions.

Justifying Claims for Model Generality. Fitting a model to data yields a set of equations. The parameter values in the equations optimize the fit of the model in the specific data set. Optimization is influenced by the effects of chance on the pattern of covariation in the data. Optimization also may be affected by the fact that the data were sampled from a specific population. Generalization tests explore the effects of chance and population differences on model structure.

Generalization is always an issue in behavioral research (Blalock, 1982). In military research, one might ask whether the same model applies to men and women, to different ethnic groups, to different occupations, and/or to different military services. For example, does general intelligence (i.e., psychometric 'g') predict job performance equally well for all military occupations?

Generalization encompasses cross-validation and moderator analysis. Cross-validation applies a model developed in a sample drawn from a particular population to a different sample from the same population. Moderator analysis compares models across samples drawn from different populations and/or situations. In both cases, the question is whether the model varies substantially from one sample to another.

Browne (2000) provides an overview of cross-validation issues for different types of analysis. His review describes model development as consisting of calibration and validation phases. However, current

guidelines use the term "validity" to refer to the appropriateness of the interpretations of test scores (APA, 1985). By extension, model validity would refer to the appropriateness of the interpretations of model parameters. This aspect of modeling can be realigned by characterizing the examination of sources of parameter variation as generalization tests. Instead of linking parameter variation to model interpretation (i.e., validity), the realignment emphasizes the legitimate scope of application of the model. This shift highlights the affinity between parameter variation and Cronbach et al.'s (1972) generalizability approach to test scores.

The effects of optimizing the fit of the model to a single sample can be estimated directly from results obtained in fitting the initial model. The shrunken R^2 printed out in many regression and GLM programs is the most familiar example of this approach. A number of indices to estimate the population accuracy of a model have been developed for regression (Raju, Bilgic, Edward, & Fleer, 1997). Users should be aware that some formulae estimate the population multiple correlation for the model; other formulae indicate the R^2 that would be expected when applying the model to a new sample of data from that population. Sampling variation specific to the new sample would affect performance in the latter case. Raju, Bilgic, Edward, and Fleer (1999) conducted a simulation to compare a number of widely used formulae. The formulae performed well when the sample size was at least moderately large (i.e., $N \geq 100$ or so). The expected cross-validation index (ECVI; Browne & Cudeck, 1989) is the analogous SEM index.

Equivalent Models. Principled argument is most productive when it compares competing models. Unfortunately, model comparison is not the norm in behavioral research (Katzko, 2002). As a result, behavioral modeling is affected by confirmation bias and insensitivity to the existence of equivalent models.

Confirmation bias is a prejudice in favor of the model under evaluation (MacCallum & Austin, 2000). Symptoms of bias include overly positive evaluations of model fit and a "… routine reluctance to consider alternative explanations of the data" (p. 213). MacCallum and Austin (2000) recommend using strategies that provide for examination of alternative models, including a priori specification of multiple models. Based on a review of recent modeling literature, these authors suggest that this approach is followed about half of the time.

A search for alternative models is likely to identify equivalent models. Two models are equivalent if they are of equal complexity and fit the data equally well. Models have equal complexity if they have the same number of parameters. MacCallum, Wegener, Ueltino, and Fabrigar (1993) found equivalent models in 46 of 53 studies they examined. The median number of equivalent models was between 12 and 21, depending on the research area. The model differences were not trivial from a theoretical perspective. Many alternative models had

very different substantive interpretations than the model adopted in the original study.

MacCallum et al.'s (1993) review understates the magnitude of the problem. That review used methods developed by Stelzl (1986) and extended by Lee and Hershberger (1990) to identify equivalent models. Raykov and Penev (1999) recently showed that those approaches are special cases of more general conditions for model equivalence. A search for all models that satisfy these general conditions would be expected to increase the MacCallum et al. (1993) estimate of the number of equivalent models per study.

Models that are literally equivalent should not be the only concern when attempting to avoid confirmation bias. Other models may fit the data nearly as well as the best model(s). The population interpretations of some SEM indices (e.g., RMSEA, ECVI) make it clear that a sample yields an estimate of the fit between the model and the data. The true population value of the GFI is most likely to fall in the range defined by the CI. Other models that are not literally equivalent to the current model will have GFI values that fall within the CI. These models should be considered along with any literally equivalent model(s). Special attention should be given to models that fit nearly as well even with fewer parameters than the current sample-optimal model. A trade-off between model accuracy and the number of parameters is the heart of the parsimony issue raised by Mulaik et al. (1989).

The statistical toolbox includes search methods to identify equivalent models. Some regression programs offer an "all possible subsets" routine. This method considers all possible combinations of the available predictors within limits set by the researcher. For example, models might be restricted to combining no more than five of eight available predictors. Large numbers of models are fitted to the data even with these restrictions. It often will be the case that several models offer comparable explanatory power. Mallows's (1973) *Cp* is a statistic that provides a parsimony index for regression. *Cp* can be used to choose between alternatives (see Draper & Smith, 1998).

The model search problem is more complex in SEM. The TETRAD program (Glymour, Scheines, Spirtes & Kelly, 1987; Spirtes, Glymour, & Scheines, 1993) provides tools that permit the computer to search for alternative models. The current version of the program permits the research to specify constraints on the search in terms of background knowledge. The background knowledge may include information about whether the population SEM includes latent traits or correlated errors, the time ordering of the variables, any established causal relationships, and causal relationships that are known not to hold in the population (Scheines, Spirtes, Glymour, Meek, & Richardson, 1998). Scheines et al. (1998) describe the basic rationale for their approach and its implementation in TETRAD II in a special issue of *Multivariate Behavioral Research,* which includes commentary. The initial TETRAD approach was compared with other search tools in a special issue of

Sociological Methods and Research (Spirtes, Scheines, & Glymour, 1990), also with attendant commentary.

The future may see tools such as TETRAD combined with developments such as Raykov and Penev's (1999) delineation of general conditions for identifying equivalent models. Separately or in combination, these tools make it possible to explore the problem of specifying equivalent models more systematically. Constructive applications of these tools could address limitations of existing research (Bentler & Dudgeon, 1996; MacCallum & Austin, 2000) to bring practice in line with recent recommendations for the proper conduct and reporting of SEMs (Boomsma, 2000; McDonald & Ho, 2002).

Causal Interpretations. Model construction, appraisal, and amendment yield one or more sets of equations. Each set represents a plausible alternative model. The sets of equations often are rendered visually as path diagrams that include unidirectional arrows representing hypothesized causal effects. Thus, the mathematical statements are routinely given causal interpretations despite cautions against this practice (Breckler, 1990; Roesch, 1999). These interpretations should be sensitive to two challenges that are related to causal inference.

Incomplete models are one source of concern. Model parameters often are interpreted as indicating the amount of change in the dependent variable that would be observed if a predictor were changed by one unit. This interpretation will err if the parameter estimate is biased. Any omitted variable produces bias if it has a causal influence on a dependent variable and is correlated with one or more predictors in the model (James et al., 1982). The extreme case is a spurious relationship. A spurious relationship arises when omitted variables are the entire basis for the association between a model predictor and a dependent variable (Kenny, 1979). James et al. (1982) discuss methods of reducing the risk of omitted variable bias.

Philosophical issues remain even after omitted variable bias has been ruled out. The general problem can be illustrated by considering the interpretation of results from a true experiment. In this case, it is impossible to directly demonstrate a causal effect on an individual. This demonstration would require observing the person as he or she would be after receiving the treatment *and* as he or she would be without the treatment. Only one of these two conditions can actually be observed, so a causal effect cannot be established for any given individual. However, in a true experiment, it is possible to estimate the average treatment effect. This parameter is an unbiased estimate of the average of unit effects. Sobel (1996, 2000) discusses these issues in greater detail. In the context of typical behavioral modeling efforts, the advice of the American Psychological Association Task Force on Statistical Inference should be kept in mind: "… especially when formulating causal questions from non-randomized data, the underlying assumptions needed to justify any causal conclusions

should be carefully and explicitly argued …" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 600).

Graph theory provides tools to address causality in connection with observational data (Glymour et al., 1987; Pearl, 1998; Spirtes et al., 1993). This approach represents the measurement and path models in an SEM as directed graphs. The directed graph includes the familiar arrows from SEMs as hypothesized causal effects. The directed graph can have testable implications such as disappearing partial correlations and TETRAD equations (Glymour et al., 1987). Determining whether the implied equations hold in the data tests the plausibility of the model. This approach cannot prove that any given model is correct, but it can rule out some competing models (Glymour et al., 1987).

HLM, CLDV, and LGCA Models

The SEM evaluation issues also apply to HLM, CLDV, and LGCA models discussed previously. Recognizing this, there appear to be opportunities to expand on current practice to obtain more complete model assessments. For example, each approach to modeling produces residuals that can be evaluated. However, standard analysis packages may not include methods of identifying influential data points. Preliminary regression analysis can serve this purpose (Raudenbush & Bryk, 2002). The GFIs from SEM can be applied to other procedures that yield χ^2 values as indicators of model fit. Thus, both the PRE and misfit indices could be applied to other areas of study. Some movement in this direction is already taking place. For example, it has been suggested that the explanatory power of models can be expressed in terms of the proportion of the null model χ^2 explained by a substantive model (Agresti, 1996; Long, 1997). Attention also has been given to examining residuals (Long, 1997).

Despite suggestions to the contrary, the analysis of categorical variables currently emphasizes significance testing. The problem of sparse data (i.e., many empty or nearly empty cells in a cross-classification) is a contributing factor (Bartholomew & Tzamourani, 1999; Collins, Fidler, Wugalter, & Long, 1993; Langeheine, Pannekoek, & van de Pol, 1996). These cells can bias the observed χ^2 upward. Collapsing cells is one means of reducing this problem, but this approach discards some of the information in the data. Bootstrap methods (cf., Efron, 1982) that avoid this loss are the recommended means of generating probability distributions for choosing between models.

Model Evaluation and MAGIC

Model evaluation is critical to principled argument. NHST provides a weak, often misleading, basis for model evaluation. Movement toward SST is desirable. Meta-analysis can facilitate movement toward SST if ES or the information required for computing ES is reported consistently. However, significance testing arguably is a

weak model appraisal tool with limited applications. Increased use of other indicators of model adequacy can be expected in the future. Movement away from purely statistical summaries toward common language indicators of ES and predictive power would be constructive, but changes in this area are likely to lag behind changes in statistical practices.

Changes in these traditional practices are likely to come slowly until alternative methods of evaluation are available. SEM appraisal practices provide one set of alternatives. This approach emphasizes a process that challenges a model by pointing out its limitations or by suggesting alternative models. Arguments based on limitations include:

- "The model is determined by influential data points and outliers."
- "The model capitalized on chance in stepwise modifications."
- "The model includes predictors that serve no useful purpose."

Arguments based on alternative models include:

- "Other models account for the data as well."
- "Other models are more parsimonious."

Any useful model must have statistically significant predictive power. Using NHST as the basis for model evaluation, therefore, represents the application of a minimum standard for model acceptance. SST is more relevant to appraisal when models progress beyond this minimum, but SST results apply to a specific model as it is currently formulated. The critical appraisal and amendment procedures are those that counter the challenges noted above. Methods that move beyond significance testing are needed to respond to those challenges.

Systematic amendment and appraisal processes help to avoid common weaknesses in the modeling process. Confirmation bias is a critical problem given the current state of the art. Modeling efforts often focus on a single model. The search for alternative models is frequently limited to adding parameters to or deleting parameters from a base model. The typical result is a final model that differs trivially from the initial model. Indeed, the modifications introduced may be no more than the effects of chance unless special steps are taken to allow for the number of significance tests involved in the modification process. The GFI or PRE for the model is likely to be interpreted optimistically. The fact that subjectively plausible ex post facto explanations can be offered for the structure of the current model may be taken as evidence of its credibility. This practice is questionable in light of Armstrong and Soelberg's (1968) demonstration that models produced by random data can be given plausible interpretations. These two model appraisal tendencies make it likely that equivalent models will be neglected. Models implied by alternative research paradigms are likely to be ignored completely

(Katzko, 2002). Careful attention to these issues in model appraisal and amendment can substantially strengthen current practices and promote principled argument.

The amendment and appraisal process directs attention to the interplay of different elements of MAGIC. These components of the model construction process emphasize articulation and credibility. Articulation is critical in defining alternative models and justifying modifications to existing models. Showing that some explanations for data are less plausible than others enhances the credibility of the better models. The plausibility of a model is increased by adjusting the weight given to magnitude estimates to allow for chance sampling effects and the effects of outlier/ influential data points. The adjusted magnitude estimates then can be weighed against other criteria (e.g., parsimony). Improved model appraisal practices are likely to reduce the initial interest value of a model. Sound practice will highlight the existence of equivalent models, the potential for capitalization on chance, and so forth. Initial assessments must weigh these facts against any novel or intellectually intriguing element(s) in the new model. Interest in the model will grow if it meets the challenges of the appraisal process.

The current state of the art poses a challenge. The power of statistical tools for fitting and refining models is increasing. This power can be used to sharpen the process of principled argument. Stronger arguments will be provided if applications of the tools are appropriately sensitive to concerns such as outliers, influential data points, the risk of capitalizing on chance when performing multiple significance tests, and the distinction between statistical significance and explanatory power. The models produced by applying those tools should be sensitive to the need for caution in causal inferences and to the likelihood that equivalent models may exist. The challenge arises because rigorous incorporation of each of these elements into model construction is not a matter of habit for most researchers. In fact, careful implementation of these desiderata requires substantial time and effort. Changing these practices can be difficult even for highly motivated investigators. However, the alternative is an increased risk of "garbage in, garbage out" behavioral models. Even if the principled argument process ultimately sorts the good from the bad, the sorting process will be far more efficient if each study is as strong as possible.

Help is available for the overwhelmed investigator. Recent recommendations for sound statistical practices (Wilkinson & the Task Force on Statistical Inference, 1999) and modeling (Boomsma, 2000; McDonald & Ho, 2002) point to the most important tools to support improved modeling. These articles could be abstracted to provide checklists that will help ensure proper attention to the appraisal problems noted here are properly addressed. Implementing those recommendations consistently will make the model construction process more challenging to the theorist and to the data analyst. The effort will be repaid by gains in model accuracy and credibility and

increased persuasiveness in the argument process. These steps turn the process into the principled argument needed to generate reliable knowledge.

Searching for New Perspectives

Even the most conscientious application of the methods described in the preceding section may not produce a satisfactory model of behavior. The range of models that can be considered is limited by the variables that are available for inclusion in the models. The range may be limited by a commitment to a given theoretical framework. In fact, research often employs paradigms that cannot be compared; the result is behavioral science encompassing several explanations that are treated as equivalent yet mutually exclusive accounts (Katzko, 2002). If each model has adherents, parallel explorations of alternative models can generate a range of useful insights. However, parallel research programs are more likely to divide the research community than to yield a consensus. From the perspective of this chapter, consensus is a necessary component of reliable knowledge (Ziman, 1978). If consensus cannot be reached, two or more different predictions could be made for the same event. In such a case, additional work is needed to determine which prediction is correct. This uncertainty can be resolved in four ways. First, one paradigm can be adopted as correct and the others discarded. Second, the paradigms can be shown to be different methods of operationalizing the same construct(s). The paradigms now become auxiliary models that demonstrate convergence of methods. Third, the paradigms can be combined to provide a more complete model. In many instances, this step will be necessary to replace models that merely provide statistically significant prediction with a model that provides a high level of predictive accuracy. Fourth, boundary conditions can be defined that determine when each paradigm is relevant to behavior. Given these alternatives, the isolated study of individual paradigms obviously can be constructive. However, research will not yield reliable knowledge as defined by Ziman (1978) until the paradigms are considered jointly. Direct comparisons are fundamental to deciding whether the explanations provided by difference models are competitive or complementary. This statement is true no matter how elegant the formal statements and tests of different models may be.

Modeling can reach an impasse despite the serious pursuit of the comparison, contrast, and integration of different paradigms. The integration may produce an overarching paradigm that includes the best elements of all available alternatives. This super paradigm still may not adequately account for behavior of interest. The principled argument process can grind to a halt if there is no method of introducing new perspectives. Qualitative research methods and exploratory data analysis (EDA) are tools for identifying new perspectives.

Qualitative research and EDA have a common core. Both approaches search for patterns in data. This common element introduces a

potential problem. Human beings are very good at perceiving patterns (Gould, 2002). The perceived patterns are translated easily into plausible stories of causal events. However, those stories may exclude key facts to conform to an iconic form (Gould, 2002; Miles & Huberman, 1994). Thus, human interpretive tendencies can work against the search for a better understanding of behavior. The search for patterns must include mechanisms to protect against this possibility. The need for an open mind is one underlying theme of the discussion of methods of searching for new perspectives that follows below. The value of checks and balances in the search is another theme. Properly combined, these elements make qualitative research and EDA constructive tools for exploring blind spots that limit the value of behavioral models.

Qualitative Research

Statistical models are mathematical abstractions that frequently are interpreted as descriptions of causal processes. The formal statements of these models appear to be definitive. A neat set of equations with specific parameter values replaces the original data. This form of presentation makes it easy to forget that the parameter values are only sample estimates, that all of the equations include an error component, and that latent variables are involved. The risk of producing nonsense is substantial if statistical models are not subjected to serious tests based on other methods.

General Approach

Qualitative research provides methods that can be used to generate initial models or to subject existing models to additional testing. Qualitative research covers a wide range of activities (Denzin & Lincoln, 1994). The focus of these activities is the identification of patterns in a set of observations. The observations may be recorded in notes made by an observer, in written material produced by the subjects being studied, or in other forms. Matrices and graphs are among the tools commonly used to identify patterns (Miles & Huberman, 1994).

Qualitative methods require a suitable database. Observations must be made and entered into databases, usually as text. The text must be annotated with observer judgments to identify critical points and link them to specific sections of material. The coding process may identify ambiguous code categories, important events that do not fit within the coding scheme, or other anomalies. In such cases, the investigator must amend the existing process and review the material again. Once the coding process is complete, the investigator must search through a large volume of material to identify specific instances of hypothesized associations. The data must then be abstracted to identify patterns that can be used to organize the findings (Miles & Huberman, 1994). After the pattern has been established, the data may be reviewed yet again to determine whether anomalies represent coding errors. The investigator then may review

the material still one more time to ensure that all events that fit within the coding scheme have been identified. Finally, the search might be followed by additional searches to test the internal logic of the existing coding scheme, evaluate tentative inferences drawn from that scheme, and identify competing explanations (Miles & Huberman, 1994). This general procedure has been facilitated in recent years by the development of a number of computer programs to aid in the process (cf., Dohan & Sanchez-Janowski, 1998; Miles & Huberman, 1994). As yet, there is no single best program or "killer application" (Dohan & Sanchez-Janowski, 1998). The basic methods of qualitative analysis have made it difficult to increase sample sizes in the past.

Search Methods

The search for patterns in qualitative data matrices can involve a variety of heuristics. Miles and Huberman (1994, pp. 245-277) draw a distinction between strategies that generate meaning and strategies that test or confirm findings. Tactics for generating meaning include (1) noting patterns and themes, (2) seeing plausibility, (3) clustering, (4) making metaphors, (5) counting exemplars, (6) making contrasts and comparisons, (7) partitioning variables, (8) subsuming particulars into general categories, (9) factoring, (10) noting (qualitative) relations between variables, and (11) finding intervening variables. The products of these tactics then are used to build a logical chain of evidence and to make conceptual or theoretical sense of the data.

Qualitative research is sensitive to the potential for biases such as perceiving events as more patterned than they actually are (Miles & Huberman, 1994, p. 263). Good qualitative research includes checks to reduce the risk of bias. Tests include checks for (1) data representativeness, (2) researcher effects, (3) methods effects, and (4) data weighting effects. Tactics for detecting points where the pattern breaks down include (5) searching for outliers, (6) examining extreme cases within the pattern, (7) reviewing surprising events, and (8) searching for data that are contrary to the pattern. Explanations are tested by (9) making if-then tests, (10) ruling out spurious relations, (11) replicating key findings, (12) checking rival explanations, and (13) getting feedback from informants.

Formal Analysis

The preceding lists of exploratory and confirmatory tactics provide a rough general picture of the qualitative research approach. The underlying logic of translating observations into theoretical statements is similar to that applied in quantitative analysis. This similarity is even more pronounced when qualitative researchers undertake formal qualitative analyses. Formal qualitative analysis techniques do not yield predictive equations, but do impose specific restrictions on the organization and interpretation of data (Griffin & Ragin, 1994).

Formal qualitative analysis techniques commonly focus on categorical variables. Models are constructed to explain why cases fall into particular categories for one of the variables. The explanatory variables in a study define a large matrix in which each cell represents a different combination of categories. When all the cases in a cell come from a single category of the criterion variable, the combination of attributes defining the cell comprise a set of conditions that are sufficient to produce a case. The simplest explanatory model results when two conditions are met. First, all of the cases in each criterion category fall in a single explanatory cell. Second, the explanatory cell is different for each criterion category. When cases from a single criterion category are found in more than one cell, more than one causal process may precede the same end state. A method such as qualitative comparative analysis (QCA; cf., Ragin, 1987) can be used to formally determine which explanatory variables actually are needed to account for membership in each criterion category.

Qualitative Causal Models

The models generated by qualitative research differ from statistical models in two important respects. Qualitative research models are based on formal logic. Causal models are formulated in terms of necessary and sufficient conditions. All cases demonstrating a specific profile of explanatory variables are expected to be members of the same outcome category. If an observation with a particular explanatory profile is not a member of the predicted outcome category, the data are reviewed to identify errors in coding. If the coding is correct, a search for additional predictors may be initiated. This approach contrasts with statistical models such as discriminant function or loglinear analyses. Those methods would estimate a set of probabilities representing the likelihood that the case should be classified as a member of each outcome category. The case then would be assigned to the category with the highest probability. Thus, qualitative analyses strive for a definite assignment of each case while quantitative analyses assign cases to categories based on probabilities. In some cases, statistical models can produce roughly equal probabilities for membership in two or more categories. The associated uncertainty is one difference between the two approaches.

The explanatory significance of predictor variables also differs between qualitative and quantitative analysis models. Statistical models focus primarily on additive effects of predictors. Most models therefore consist of linear weighted sums of the predictors. For each observation, the probability of category membership is increased or decreased to some extent based on the predictor score. The probability estimate is modified regardless of the values of other predictors. In the qualitative approach, none of the predictors has an isolated effect. The import of each predictor is contingent on the values of other predictors because a case is assigned to a particular category only when the overall profile of explanatory variables justifies that classification. The contingent nature of the

relationship between explanatory variables and category membership would imply an interaction in a statistical model. A qualitative model involving several predictors, therefore, might be equivalent to a statistical model involving higher order interactions. A qualitative model with even a modest number of predictors therefore implies a level of complexity of interplay among predictors seldom found in statistical models. From the qualitative perspective, the complexity is justified by the assumption that causal processes that determine category membership represent the interplay of a number of factors. A theoretical account of the evidence must spell out the contingencies in this interplay. Interpretations of statistical models are less likely to assume that the set of predictors in the model define an integrated causal process. Instead, those models are likely to interpret category membership as the product of the independent operation of a number of independent causal processes. Qualitative analyses, therefore, may be especially useful in stimulating thought about the interplay of causal variables.

Qualitative Research and MAGIC

Qualitative research emphasizes two elements of the MAGIC model that are likely to receive less attention in quantitative studies. The qualitative approach certainly emphasizes the articulation of causal processes and their links to actual behavior and events. The qualitative approach also links models to real entities and events. This linkage is likely to make the results more interesting to consumers of the model.

Statistical models often embody very sketchy causal assertions. A sketchy description of a plausible rationale is given and an appropriate arrow is inserted into a causal graph. One set of causal arrows is preferred if it reproduces aggregated observations better than another set. This avenue of study can be pursued without ever subjecting the initial causal assertions to close scrutiny. For example, it may never be determined whether the assertions are plausible for a single case considered in isolation.

By contrast, qualitative analysis can subject causal assertions to closer scrutiny. Abstract traits are replaced with specific events that often can be located in temporal sequences. Serious consideration may be given to alternative causal paths without the necessity of choosing a single alternative. For example, if QCA produces more than one cell of "cases," the result implies the existence of alternative causal models. These models might be represented in an SEM as different pathways once identified, but the key problem of identifying alternative causal patterns would be more difficult to solve in the usual quantitative analysis. Insensitivity to the existence of alternative causal models is a weakness of current practice in statistical modeling.

Qualitative research can increase the interest value of models. Statistical models in the behavioral sciences are of interest

primarily to narrow research communities. Economics models are an obvious exception to this statement. In other areas, the linkage between latent traits and specific types of behavior may be too vague to interest potential consumers (e.g., policy-makers, clinicians, managers, military leaders). Abstract variables are of interest to these audiences only when they map onto the decision terrain faced by the user. Examples from qualitative research could help to define this relationship more clearly.

Exploratory Data Analysis (EDA)

EDA (Tukey, 1977) shares a core element with qualitative research. Both approaches are concerned with exploiting the richness of the data. This concern drives the view that models should reflect observations made by the investigator after a period of intensive interaction with the objects of study. Both approaches share a concern that routine application of statistical computer algorithms will obscure important aspects of the data. Thus, both EDA and qualitative research emphasize cyclical evaluation of models. Each cycle involves a sequence of identifying patterns in the data, developing hypotheses to account for those patterns, followed by testing and revising the hypotheses. The revised hypotheses then are the basis for the next cycle. The cycle is repeated until an acceptable representation of the data is obtained. EDA and qualitative research differ in that the former typically applies the observe-test-revise-test-repeat cycle to quantitative data rather than nominal categorical data.

A typical EDA sequence might be as follows. A series of graphic displays is examined to identify general patterns in the data. An initial mathematical model is formulated as a first attempt to capture the pattern. The data are analyzed to estimate parameter values for the initial model and to compute differences between the predicted and observed values. A second round of graphic displays examines the residuals from the initial model to identify areas of misfit between the model and the data. A revised model is formulated to account for the residuals and is fitted to the data. The cycle is repeated until a good representation of the data can be achieved.

The EDA approach is more a frame of mind than a unique analytic method. Any of the steps described in the preceding paragraph could be included in a standard statistical analysis. Behrens (1997) summarizes the key elements of the EDA frame of mind as:

- Understand the context well enough to make informed decisions given theory and prior research findings (p. 135).
- Use graphic representations of the data to guide analysis decisions by looking at the actual pattern of data (p. 135).

- Develop models iteratively from tentative model specification followed by residuals assessment (p. 139).
- Use robust/resistant procedures to minimize the influence of distributional assumptions (p. 143).
- Attend to outliers not merely as indications of problems in the research process, but as potential indicators of anomalous phenomena that require explanation (p. 144).
- Re-express the original scales when doing so will facilitate interpretation, promote symmetry, stabilize the spread of values within groups in the analysis, or promote straight line relationships (p. 145).

Behrens (1997) provides greater detail on the preceding points with references to original sources that explore these various issues in depth. A full treatment of these methods is not possible here, but Behrens's (1997) general guidelines are directly related to issues discussed earlier in this chapter. Understanding context is related to the idea that one should not suspend judgment when analyzing data. When prior findings contribute to context, these admonitions share the spirit of strong significance tests because prior findings replace the null hypothesis as a research field matures. The admonition to develop models iteratively is implicitly related to parsimony because it is based on fitting a simple model to data. New parameters are added only if they predict the residuals. Stepwise regression, EFA, and other analyses begin with simple models then extend them if adding more predictors or more factors will provide a better account of the variance. These procedures provide iterative models, but the methods are constrained by statistical criteria (i.e., maximize the variance explained) rather than theoretical or empirical context and the judgment of the researcher. The emphases on robust procedures and outliers directs attention to the need to develop models that accurately predict behavior in most of the people most of the time. Both of these elements of EDA could be useful to identify exceptional groups of observations. If there are no obvious errors in the data, these groups might become the basis for hypotheses that could be tested later using taxometric methods. Finally, the emphasis on interpretation is a reminder that statistical summaries are not the endpoint for analysis. The data must be interpreted in ways that link them to actual behavior and to theory.

State of the Art

Several factors make it likely that there will be dramatic improvements in behavioral models in the near future. First, the development of computer hardware and software has reduced barriers to incorporating advanced procedures into research. Today's programs routinely include simple methods of specifying models for analysis. Examples are drop-down menus and graphic interfaces. Database translation programs make it possible to format data in almost any familiar form and import it into a new program. Thus, it is no longer necessary to master a complex computer syntax that is specific to a

particular computer program with a limited range of analytic functions. For a modest cost, every researcher can have ready access to each type of analysis discussed in this chapter. In fact, researchers in large organizations are likely to have access to several different programs that implement the most widely used methods.

A reduced risk of "garbage in, garbage out" analysis is a second positive factor. The increased integration of different analysis procedures under the heading of GLM or EM maximum likelihood methods makes it clear that different models are manifestations of general principles. Long (1997) provides a fundamental expression of this point through his observation that linking functions translate different CLDV analyses into familiar linear regression models. Long further notes that techniques learned in the more familiar linear regression context apply to procedures such as logit, probit, and logistic regression analyses. Analogous situations can be identified in SEM and other procedures. Recognizing and capitalizing on these similarities reduces the learning curve required for effective use of new procedures.

There is movement toward confirmatory methods. Confirmatory methods encourage explicit theoretical statements by making it possible to impose constraints on a model. This aspect of analysis is not new. For example, many researchers have conducted analyses that forced the entry of predictors into a regression equation. However, the range of analyses that can be conducted with constraints that specify precise values for model parameters now extends to factor analysis (i.e., CFA), cluster analysis (i.e., EM mixture analysis), and substantive models (e.g., SEM). Meta-analysis provides tools for developing parameter estimates based on the cumulative research record. The process of thinking through the potential constraints encourages better articulation of the relationship between the model and theory. At the same time, a good fit between a highly constrained model and the data provides the convergence between predictions and evidence that indicates a strong theory.

The identification of important blind spots in traditional research practices is another positive development. The most critical blind spot is the tendency toward confirmation bias. The demonstrable existence of equivalent models is the strongest argument for giving careful attention to this problem. However, acknowledging confirmation bias also directs attention to other important problems. Parsimony, an accepted desideratum for sound theories, comes into view when it is recognized that nearly equivalent models may exist that involve fewer parameters. Recognition of confirmation bias can also lead to more frequent comparisons of models based on different research paradigms (Katzko, 2002). One intriguing issue here is that syndrome models, which generate an overall model by simply collecting and enumerating specific paradigms, may prove defensible on further analysis. Conceivably, this approach is the modeling equivalent of taxometric definitions of mental health syndromes. However, as Meehl

(1992) has pointed out, the existence of a typology is a testable hypothesis, not something to be established by fiat. Where multiple models or paradigms are not currently available, tools for searching for alternative models are available (e.g., qualitative analysis, EDA, TETRAD). Spirtes, Richardson, Meek, Scheines, and Glymour (1998) argue that a serious search for alternative models should be undertaken prior to conducting any analysis.

Recent publication guidelines for statistical practices should encourage improved modeling practices. The availability of these guidelines indicates that practice has matured to produce at least a broad general consensus on methods. The consensus includes recommendations on the general problems of statistical inference (Wilkinson & the Task Force on Statistical Inference, 1999) and recommendations for standard reporting in SEM (Boomsma, 2000; McDonald & Ho, 2002). The SEM guidelines may be particularly useful for modeling. The general steps that are outlined can be adapted to almost any analysis, particularly those involving the imposition of constraints on model parameters and structure. Diagnostic tools are available at various steps in the process to assess potential weaknesses of existing models. These tools include methods of identifying outlier/influential data points and searching for alternative models. Systematic application of these tools will reduce the likelihood that models will be affected by blind spots in the conceptual model under investigation or quirky elements of the data being analyzed. Applications that embody the test-and-revise spirit of EDA are likely to be particularly fruitful.

The state of the art is itself an example of principled argument. Progress has been made in some areas, but consensus has not been reached on all aspects of modeling. Methods of appraising and amending models are in a state of flux. Significance testing is becoming less important in SEM, but it continues to be the primary tool for model assessment in other areas (e.g., CLDV analyses). Even in the SEM domain, consensus is only qualitative in some areas. For example, no consensus has been reached regarding the best GFI to use. Hu and Bentler's (1998, 1999) two-indicator approach probably approximates the current consensus in this area with SRMR and RMSEA as a workable combination. These indices reflect the misfit and PRE of the model, respectively, and appear to be sensitive to mistakes in both the measurement model and path model. Uncertainty in this area illustrates an issue that is likely to be important in SEM in the near future. The ongoing controversy over significance testing directs attention to the potential use of multiple criteria in other areas. Given multiple criteria, it is reasonable to expect future work to address the problem of how best to combine alternative criteria. At present, the issue of selecting and weighting indicators of model adequacy is a judgment call for the researcher. At the same time, there is clear evidence of movement away from relying on significance testing as the primary method of model evaluation.

It is not clear at this time whether the PRE approach should be extended to all types of models. For example, should this criterion be applied in the study of CLDV? The research tradition in areas of study using these tools has emphasized significance testing rather than incremental fit as the primary basis for choosing between models. Procedures such as Kaplan's (1990a, 1990b) combination of modification indices and expected parameter change provide an alternative perspective on post hoc model modification. However, investigators must be sensitive to the risk of producing a complex model that merely capitalizes on chance (Green et al., 2001; Steiger, 1990). Responses to Kaplan's (1990a) suggestions included the recommendation that modifications should not be introduced without adequate theoretical justification (Bollen, 1990). Kaplan (1990b) concurred with this recommendation, but even this criterion may be inadequate. Steiger (1990) posed the question, "What percentage of researchers would ever find themselves unable to think up a *theoretical* justification for freeing a parameter?" (p. 175, italics in the original). Note that this question was posed in the context of post hoc modifications rather than a priori specification. Even with such justification, modifications should be examined in a new sample of data to verify that they are productive. In connection with this point, Steiger (1990, p. 176) also noted emphatically, "*An ounce of replication is worth a ton of inferential statistics*" (italics in the original). Increasing use of replication will likely be a trend in the future. One reason is that increased use of bootstrapping and other resampling methods provides methods of pursuing this end without radically increasing the volume of data needed in the modeling process (Wilcox, 1998).

The development of a broader perspective on research programs may become a growth area in the future. Qualitative research and EDA have been examined here as potential methods of avoiding confirmation bias. Qualitative analysis can be an end in its own right, but this general approach also has the potential to stimulate the formulation of new models. QCA is interesting as a means of identifying causally relevant variables that could be incorporated into models. QCA also is a stimulus to rethinking a problem because it embodies a different concept of causation than is found in SEM, for example. EDA and TETRAD provide additional tools for using data to generate causal models. The use of these tools is important as an antidote to the confirmation bias that occurs when a moderately good fit to the data is interpreted as sufficient justification to accept a model specified at the outset. Perhaps a careful wedding of qualitative analysis to quantitative analyses would help overcome the resistance to qualitative research in some domains (e.g., psychology journals; Kidd, 2002). Explication of the limitations of standard statistical procedures as model generating tools coupled with careful demonstration of the checks and balances involved in proper causal inference from qualitative data could be critical to making a better case for combining the two approaches when constructing models. The development of methods of determining whether a qualitative model is superior to a quantitative model in some situations may be an

important topic for future work. For example, a qualitative model might identify a change of circumstances in an operational setting that signals the need to switch causal models. A model appropriate to the new circumstances would replace the quantitative causal model currently in use. The rapid growth of techniques for data mining from documents and other sources could facilitate the application of qualitative analysis to this type of problem in the near future.

This description of the state of the art is certainly incomplete. Attention has been limited largely to tools routinely used in psychology and sociology. Developments in other domains may provide tools that can be very useful. For example, time-series analysis was developed primarily in the realm of econometrics. The absence of any consideration of how to translate statistical models into dynamic simulations for forecasting is another potential problem. At a minimum, this translation will have to include distributions of parameter values in addition to functional relationships between variables. Distributions must be considered to generate samples of entities (i.e., individuals, groups) that will perform the actions that lead to the forecast. Techniques such as HLM are useful in this regard because they provide estimates of variances. However, the actual process of translating those estimates into simulations may be more difficult than it first appears.

Finally, the potential for different categories of models has been referred to previously only in passing. Models such as game theory may not translate readily into quantitative terms. The problem of how to construct hybrid models that integrate quantitative and qualitative differences in dynamic representations may be a major challenge for the future.

Computer Programs and Specific Implementations

This chapter has not discussed specific software programs for implementing state-of-the-art analyses. However, almost every procedure referred to in this chapter can be implemented using several statistical packages. The programs often reduce the problem of specifying a model to simple activities such as drawing a picture or filling in boxes on a pop-up computer menu. The basic problem of how to implement these advanced methods therefore reduces to choosing an appropriate program and specifying the model of interest. The full range of analysis problems that can be addressed by these means cannot be described because computer programs are being revised and updated so rapidly. Even relatively inexperienced investigators can apply advanced methods effectively when guided by recent recommendations regarding statistical practices (e.g., Behrens, 1997; Boomsma, 2000; McDonald & Ho, 2002; Wilkinson & the Task Force on Statistical Inference, 1999).

An informal survey of advanced data analysis packages suggests five trends in the development of computer analysis tools. First, newer programs emphasize model testing and comparison. Program input

includes equations that define a model to be fitted to the data. The program then estimates parameter values. Program output typically includes an overall measure of fit between the model and the data (e.g., a maximum likelihood χ^2). Second, newer programs capitalize on the fact that many nonlinear models can be transformed to linear models (cf., Long, 1997). Thus, a single program fits models that appropriate expressions for continuous and discrete variables. Third, programs increasingly accommodate different combinations of continuous and discrete variables. Either type of variable may appear as a predictor or a dependent variable in equation form. Even latent variables can be continuous or discrete (Magidson & Vermunt, 2001, 2002). Fourth, programs are more likely to provide simple methods of cross-validating models. Some programs provide options that automatically divide the data into calibration and cross-validation samples. Fifth, the range of graphic display methods is increasing. This trend makes it easier to apply EDA principles during data analysis.

These trends provide better tools for solving the difficult problem of moving from verbal statements to mathematical models evaluated by tests of fit to the data rather than by null hypothesis tests. Wider use of these tools will surely help to define consensible facts (Ziman, 1978) by clearly articulating the links between data measurements and constructs and imposing theoretically derived values on the data. Stronger consensus should also be fostered by the direct comparison of alternative models conducted in the context of metatheoretical criteria for model choice (e.g., parsimony).

Specific programs suitable for addressing a particular problem can be identified several ways. A review of the related research literature can identify programs used in prior work. Specialized journals (e.g., *Structural Equation Modeling*) often provide examples of different programs. Methodological and statistical journals review books and computer programs that describe specific programs and often contrast a given product with competitors (e.g., *Journal of the American Statistical Association*, *British Journal of Mathematical Psychology*, and *Educational and Psychological Measurement*). Internet searches can identify programs for general types of analysis. For example, at the time of this writing, a search for "latent class analysis" identified a site listing 15 computer programs that would perform this procedure. Similarly, a search for "cluster analysis" identified several programs that implement the multivariate mixture approach described by Fraley and Raftery (2002).

Guidance on specific analysis problems is available in many cases. Program documentation now routinely supplements written manuals with computerized tutorials and application examples. Textbooks that describe the underlying statistical models, the development of the analysis methods, and application examples are available and, in some cases, specifically linked to particular programs (e.g., McCutcheon, 1987; Raudenbush & Bryk, 2002; Waller &

Meehl, 1998). Texts on general topics such as SEM are widely available, but care is needed when choosing a text to ensure that it covers critical issues (Steiger, 2001). Journal articles often include appendices giving the command syntax for specific methods or models. Such appendices are common, for instance, in *Structural Equation Modeling* and *Psychological Methods* articles. Internet sites for user groups include bulletin boards for seeking expert advice on specific problems (e.g., SEMNET). These resources are helping to break down barriers to the use of modern analysis procedures, thereby providing tools for more focused tests of hypotheses.

The range of analytical options is daunting and may even be intimidating. Most people will experience a natural tendency to cling to familiar methods that provide reasonably sound answers to their questions and permit work to move forward in an orderly fashion. This reaction should be tempered by the fact that the temptation is shared with most other colleagues. At present, the average researcher is not fully prepared to exploit all analytical opportunities (Tinsley & Brown, 2000b), but this situation is neither new nor an insurmountable impediment to progress. As Berk (1997, p. xxi) notes in his introduction to Long's (1997) description of CLDV, "For most of the procedures discussed…there exist statistical routines in all of the major statistical packages. This is both a blessing and a curse. The blessing is that minimal computer skills are required. The curse is that minimal computer skills are required. Right answers and wrong answers are easy to obtain." However, if researchers remember that "*No statistical procedure should be treated as a mechanical truth generator*" (Meehl, 1992, p. 152, italics in the original), progress toward Ziman's (1978) goal of consensus should be more rapid in the future than it has been in the past. In the end, investigators who invest the time to familiarize themselves with newer techniques will be repaid by substantial gains in their ability to derive more definite answers to their research questions.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129-133.

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-144). Mahwah, NJ: Erlbaum.

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: John Wiley & Sons.

Aldenderfer, M. S., & Blashfield, R. K. (1985). *Cluster analysis.* Beverly Hills, CA: Sage Publications.

Algina, J. & Moulder, B. C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement, 61*, 633-649.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411-423.

Anderson, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on Fornell and Yi. *Sociological Methods & Research, 20*(3), 321-333.

Andrews, F. M., Klem, L., Davidson, T. M., O'Malley, P. M., & Rodgers, W. L. (1981). *A guide for selecting statistical techniques for analyzing social science data* (2nd ed.). Ann Arbor, MI: Institute for Social Research.

Andrews, F. M., Klem, L., O'Malley, P. M., Rodgers, W. L., Welch, K. B., & Davidson, T. N. (1998). *Selecting statistical techniques for social science data: A guide for SAS*. Cary, NC: SAS Institute.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago: SmallWaters Corporation.

Arminger, G., Clogg, C. C., & Sobel, M. E. (Eds.). (1995). *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum Press.

Armstrong, J. S., & Soelberg, P. (1968). On the interpretation of factor analysis. *Psychological Bulletin, 70*(5), 361-364.

Barnett, V., & Lewis, T. (1994). Outliers in statistical data (3rd ed.). New York: John Wiley & Sons.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research, 27*(4), 525-546.

Beauchaine, T. P., & Beauchaine, R. J., III. (2002). A comparison of maximum covariance and *K*-means cluster analysis in classifying cases into known taxons. *Psychological Methods, 7*(2), 245-261.

Bedeian, A. G., Day, D. V., & Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: Some important reminders. *Educational and Psychological Measurement, 57*, 785-799.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131-160.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*, 563-592.

Berk, R. (1997). Series editor's introduction. In J. S. Long (Ed.), *Regression models for categorical and limited dependent variables* (pp. xx-xxi). Thousand Oaks, CA: Sage Publications.

Blalock, H. M., Jr. (1969). *Theory construction: From verbal to mathematical formulations*. Englewood Cliffs, NJ: Prentice-Hall.

Blalock, H. M., Jr. (1982). *Conceptualization and measurement in the social sciences*. Beverly Hills, CA: Sage Publications.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K. A. (1990). A comment on model evaluation and modification. *Multivariate Behavioral Research, 25*(2), 181-185.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314.

Bollen, K. A., & Ting, K.-F. (2000). A TETRAD test for causal indicators. *Psychological Methods, 5*(1), 3-22.

Boomsma, A. (2000). Reporting analysis of covariance structures. *Structural Equation Modeling, 7*(3), 461-483.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107*(2), 260-273.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108-132.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Browne, M. W., MacCallum, R. C., Kim, C-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7*(4), 403-421.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509-540.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Chatterjee, S., & Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. *Applied Psychological Measurement, 16*(3), 209-227.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cleland, C. M., Rothschild, L., & Haslam, N. (2000). Detecting latent taxa: Monte Carlo comparison of taxometric, mixture model, and clustering procedures. *Psychological Reports, 87,* 37-47.

Cohen, J. (1978). Partialed products are interactions; partialed powers are curve components. *Psychological Bulletin, 85,* 858-866.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28*(3), 375-389.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161-172.

Costa, P. T., Jr., & McCrae, R. R. (1992) *NEO-PI-R Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53,* 585-596.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37,* 553-558.

Cronbach, L. J., Gleser, g. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* NY: John Wiley and Sons.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.

Davison, M. L. (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin, 97*(1), 94-105.

Davison, M. L., & Sireci, S., G. (2000). Multidimensional scaling. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 325-353). San Diego, CA: Academic Press.

Denzin, N., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage Publications.

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology, 53*, 133-149.

Dohan, D., & Sanchez-Jankowski, M. (1998). Using computers to analyze ethnographic field data: Theoretical and practical considerations. *Annual Review of Sociology, 24*, 477-498.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley & Sons.

Duncan, C., Jones, K., & Moon, G. (1993). Do places matter: A multi-level analysis of regional variations in health-related behaviour in Britain. *Social Science and Medicine, 37*, 725-733.

Ebrahim, S. (2003). The use of numbers needed to treat derived from systematic reviews and meta-analysis. *Evaluation & the Health Professions, 24*, 152-164.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119-127.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Edward Arnold.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial. *Educational and Psychological Measurement, 61*, 517-531.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), 56-83.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 532-574.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210.

Fornell, C., & Yi, Y. (1992). Assumption of the two-step approach to latent variable modeling. *Sociological Methods & Research, 20*(3), 291-320.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611-631.

Gellatly, I. R., & Irving, P. G. (2001). Personality, autonomy, and contextual performance of managers. *Human Performance, 14*, 231-245.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Orlando, FL: Academic Press.

Gore, P. A., Jr. (2000). Cluster analysis. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 298-324). San Diego, CA: Academic Press.

Gorsuch, R. L. (1983). *Factor analysis*. Philadelphia: Saunders.

Gould, S. J. (2002). *I have landed: The end of a beginning in natural history*. New York: Harmony Books.

Green, S. B., Thompson, M. S., & Poirer, J. (2001). An adjusted Bonferroni method for elimination of parameters in specification addition searches. *Structural Equation Modeling, 8*(1), 18-39.

Greenwald, A. O., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175-183.

Griffin, L., & Ragin, C. C. (1994). Some observations on formal methods of qualitative analysis. *Sociological Methods & Research, 23*(1), 4-21.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 57,* 15-24.Harlow, L. L. (1997). Significance testing introduction and overview. In Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* (pp. 1-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harris, C. W. (1963). *Problems in measuring change.* Madison, WI: University of Wisconsin Press.

Hays, W. L. (1963). *Statistics for psychologists.* NY: Holt, Rinehart, Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction analysis of cross-classifications.* NY: John Wiley & Sons.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325-344.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315-332.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193-218.Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193-205.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park: Sage Publications.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300-307.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage Publications.

Jones, M. B. (1998). Behavioral contagion and official delinquency: Epidemic course in adolescence. *Social Biology, 45*, 134-142.

Joreskog, K. G. (1998). Interaction and nonlinear modeling: Issues and approaches. In R. E. Schumacker, & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 239-250). Mahwah, NJ: Lawrence Erlbaum Associates.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

Joreskog, K. G., & Sorbom, D. (1996). *PRELIS™ 2: User's reference guide*. Chicago: Scientific Software International.

Joreskog, K. G., & Yang, F. (1996). Non-linear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaplan, D. (1990a). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*(2), 137-155.

Kaplan, D. (1990b). A rejoinder on evaluating and modifying covariance structure models. *Multivariate Behavioral Research, 25*(2), 197-204.

Katzko, M. W. (2002). The rhetoric of psychological research and the problem of unification in psychology. *American Psychologist, 57*(4), 262-270.

Kaufman, J. D., & Dunlap, W. P. (2000). Determining the number of factors to retain: A Windows-based FORTRAN-IMSL program for parallel analysis. *Behavior Research Methods, Instruments, & Computers, 32*(3), 389-395.

Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley & Sons.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*(1), 201-210.

Keselman, J. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative to familywise and comparisonwise Type I error control. *Psychological Methods, 4*(1), 58-69.

Kidd, S. A. (2002). The role of qualitative research in psychological journals. *Psychological Methods, 7*(1), 126-138.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44*(448), 1372-1381.

Krieger, A. M., & Green, P. E. (1999). A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika, 64*(3), 341-353.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16-26.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods,* 7(2), 228-244.

Langenheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24*(4), 492-516.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research, 24*(3), 365-395.

Lee, S., & Hershberger, S. L. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25*(3), 313-334.

Lett, R. R., Hanley, J. A., & Smith, J. S. (1995). The comparison of injury severity instrument performance using likelihood ration and ROC curve analyses. *Journal of Trauma: Injury, Infection, and Critical Care, 38,* 142-148.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187-192.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (Vol. 7). Thousand Oaks, CA: Sage Publications.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113-139.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201-226.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of

capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

MacCallum, R. C., Wegener, D. T., Ueltino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*(1), 185-199.

MacCallum, R. C., Widaman, K. F., Shang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Magidson, J., & Vermunt, J. K. (2001) Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology, 31*, 223-264.

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*, 37-44.

Mallows, C. P. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*(4), 335-361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*(1), 47-70.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2*(1), 3-19.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376-390.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall.

McCutcheon, A. L. (1987). *Latent class analysis* (Vol. 64). Newbury Park, CA: Sage Publications.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

McLachlan, G. J., & Krishnan, T. (1997). The EM algorithm and extensions. New York: Wiley.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60,* 117-174.

Meehl, P. E. (1995). Bootstrap taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50*(4), 266-275.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391-423). Mahwah, NJ: Erlbaum.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553-557.

Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement, 61,*650-667.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R. Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological Testing and Psychological Assessment: A review of evidence and issues. *American Psychologist, 56,* 128-165. Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis.* Thousand Oaks, CA: Sage Publications.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for determine the number of clusters in a data set. *Psychometrika, 50,* 159-179.

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26,* 479-497.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430-445.

Mulaik, S. A., Raju, N. S., & Harshman, R. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-116). Mahwah, NJ: Erlbaum.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299-1301.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*(3), 396-402.

Overall, J. E., & Magee, K. N. (1992). Replication as a rule for determining the number of clusters in hierarchical cluster analysis. *Applied Psychological Measurement, 16*(2), 119-128.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research, 27*(2), 226-284.

Ping, R. A., Jr. (1995). A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research, 32*, 336-347.

Ping, R. A., Jr. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin, 119*(1), 166-175.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160-164.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkely, CA: University of California Press.

Raju, H. S., Bilgic, R., Edward, J. E., & Fleer, P.F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement, 21*, 291-305.

Raju, H. S., Bilgic, R., Edward, J. E., & Fleer, P.F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23*, 99-115.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846-850.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52*, 501-525.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34*(2), 199-244.

Rigdon, E. E., Schumacker, R. E., & Wothke, W. (1998). A comparative review of interaction and nonlinear modeling. In R. E. Schumacker & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum Associates.

Roesch, S. C. (1999). Modeling stress: A methodological review. *Journal of Behavioral Medicine, 22*(3), 249-269.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726-748.

Rogosa, D., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50*(2), 203-228.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage Publications.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59-82.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research.* New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology, 9*(5), 395-396.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 10,* 1276-1284.

Rounds, J., & Tracey, T. J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology, 78*(6), 875-890.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection. New York: John Wiley & Sons.*

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology, 77*(3), 336-342.

Russell, C. J., Pinto, J. K., & Bobko, P. (1991). Appropriate moderated regression and inappropriate research strategy: A demonstration of information loss due to scale coarseness. *Applied Psychological Measurement, 15*(3), 257-266.

Salmon, W. C., (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16,* 209-222.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD Project: Constraint based aids to causal model specification. *Multivariate Behavioral Research, 33*(1), 65-117.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129.

Schmidt, F. L. & Hunter, J. E. (1998). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology, 84,* 144-148.

Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interaction and nonlinear effects in structural equation modeling.* Mahwah, NJ: Erlbaum .

Seaman, M.A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110*(3), 577-586.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73-83.

Shapiro, D. (1965). *Neurotic styles.* NY: Basic Books.

Sliwinski, M. J., & Hall, C. B. (1998). Constraints on general linear slowing: A meta-analysis using hierarchical linear models with random coefficients. *Psychology and Aging, 13*(1), 164-175.

Smith, L. D., Best, L. A., Cylke, V. A., & Stubbs, D. A. (2000). Psychology without *p* values: Data analysis at the turn of the 19th century. *American Psychologist, 55,* 260-263.

Smith, L. D., Best, L. A., Stubbs, D. A., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist, 57,* 749-761.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605-632.

Sobel, M. E. (1996). An introduction to causal inference. *Sociological Methods & Research, 24*(3), 353-379.

Sobel, M. E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association, 95*(450), 647-651.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research, 27*(1998), 182-225.

Spirtes, P., Scheines, R., & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using

TETRAD II, EQS, and LISREL programs. *Sociological Methods & Research, 19*, 3-66.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96*(453), 331-338.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21*, 309-321.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95*(2), 334-344.

Strube, M. J. (1989). Evidence for the Type in Type A behavior: A taxometric analysis. *Journal of Personality and Social Psychology, 56*, 972-987.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525-534.

Thompson, B. (2002). 'Statistical,' 'Practical,' and 'Clinical': How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Tinsley, H. E. A., & Brown, S. D. (Eds.). (2000a). *Handbook of applied multivariate statistics and modeling*. San Diego, CA: Academic Press.

Tinsley, H. E. A., & Brown, S. D. (2000b). Multivariate statistics and mathematical modeling. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 1-36). San Diego, CA: Academic Press.

Tracey, T. J., & Rounds, J. (1993). Evaluating Holland's and Gati's vocational-interest models: A structural meta-analysis. *Psychological Bulletin, 113*(2), 229-246.

Trafimow, D. (2000). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110*, 526-535.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.

Vacha-Haase, T. & Ness, C. M. (1999). Statistical significance testing as it relates to practice: Use within professional psychology: Research and practice. *Professional Psychology: Research and Practice, 30,* 104-105.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321-327.

Vickers, R. R., Jr., Hervig, L. K., Wallick, M. T., & Conway, T. L. *The Marine Corps basic training experience: Correlates of platoon attrition rate differences* (Tech. Rep. 84-9). San Diego, CA: Naval Health Research Center.

von Eye, A., & Brandstadter, J. (1998). The wedge, the fork, and the chain: Modeling dependency concepts using categorical variables. *Psychological Methods, 3*(2), 169-185.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212-213.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua.* Thousand Oaks, CA: Sage Publications.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300-314.

Wilkinson, L. & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Winer, B., Brown, D., & Michels, K. (1991). *Statistical principles in experimental design.* New York: McGraw-Hill.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5,* 329-350.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*(4), 354-365.

Zedeck, S. (1971). Problems with the use of moderator variables. *Psychological Bulletin, 76,* 295-310.Ziman, J. (1978). *Reliable knowledge: An exploration of the grounds for belief in science.* London: Cambridge University Press.

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY)	2. Report Type	3. DATES COVERED (from - to)
14 Nov 03	final	01-01-03 - 11-1-03

4. TITLE AND SUBTITLE
Statistics and the Art of Model Construction

6. AUTHORS
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Office of Naval Research
800 N. Quincy St., BCT 1
Arlington, VA 22217-5660

Chief, Bureau of Medicine and Surgery
Code M2
2300 E Street N.W
Washington DC 20372-5300

5a. Contract Number: USAMMRC
5b. Grant Number: Reimb-60109
5c. Program Element 637006N
5d. Project Number: M0096
5e. Task Number: 001
Work Unit Number: 6417
5g. IRB Protocol Number: NA

8. PERFORMING ORGANIZATION REPORT NUMBER
Report 04-08

10. Sponsor/Monitor's Acronyms(s)
BuMed/ONR

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

Behavioral models should be based on reliable knowledge. Reliable knowledge is achieved when a scientific community reaches consensus on the interpretation of available evidence. When properly used and interpreted, statistical models can aid in the process of principled argument that leads to consensus. This report reviews the state of the art for the use of statistical methods in behavioral modeling. The major topics covered are the construction of measurement models to quantify behavioral constructs, the construction of path models to describe relationships between behavioral constructs, issues associated with model appraisal and amendment, and methods of searching for alternatives to or refinements of existing models. Examples of specific topics include methods of evaluating the underlying nature of constructs (i.e., continuous or categorical), methods of constructing models that combine constructs from different research domains (e.g., individual growth, individual differences, and group processes), the place of significance tests in model evaluation, and methods of searching for new insights regarding behavior. Potential military applications are illustrated. Each section concludes by considering how the issues and methods discussed contribute to the process of principled argument that is needed to ensure that behavioral models are based on reliable knowledge.

14. SUBJECT TERMS
statistical methods, model appraisal, exploratory data analysis, measurement models, path models

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGE	18a. NAME OF RESPONSIBLE PERSON
a. REPORT	b.ABSTRACT	C. THIS PAGE			Commanding Officer
UNCL	UNCL	UNCL	UNCL	79	18b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18